SCIENTIFIC COMMITTEE
NINETEENTH REGULAR SESSION

Koror, Palau
16 – 24 August 2023

Independent Review of Recent WCPO Yellowfin Tuna Assessment

WCPFC-SC19-2023/SA-WP-01

André E. Punt[1], Mark N. Maunder[2], and James N. Ianelli[3]

Comments on this ***Peer Review Report*** through the **Online Discussion Forum** (https://forum.wcpfc.int/c/sc-19/27) are due **by 31 March**.

[1] Professor of Aquatic and Fishery Sciences at the University of Washington
[2] Head of the Stock Assessment Program at the Inter-American Tropical Tuna Commission
[3] Affiliate professor at the University of Washington and a senior scientist with the Resource Ecology and Fisheries Management division of the Alaska Fisheries Science Center, NOAA

Independent review of recent WCPO yellowfin tuna assessment

André E. Punt, Mark N. Maunder, and James N. Ianelli

Executive Summary

1. The 2020 assessment of yellowfin tuna was a major undertaking and involved the use of many state-of-the art techniques. It involved several changes to the 2017 assessment, including the use of more data sources, improved diagnostics, and improved methods for parameterizing and specifying the model.
2. Major changes from the 2017 to the 2020 diagnostic models included: a) reproductive output-at-age and natural mortality-at-age were modified to use a different sex ratio-at-age vector based on updated sex-ratio-at-length data and an updated growth curve, b) the reproductive output-at-age vector was modified for the 2020 assessment based on an externally calculated reproductive output-at-length ogive, c) conditional age-at-length data were included in the assessment, as part of the diagnostic model, to better inform growth (with the form of the growth curve consequently changed), d) the tag data used in the model changed due to: i) a reduction in the tagger-related mortality resulting from changes to the how tagger effects were modelled, ii) including the Japanese tagging data; iii) inclusion of tag releases available since the last assessment, and iv) changing how tag recaptures were assigned to mixing periods, and v) including recaptures that occurred within the mixing period that had details on capture regions and dates but the recapture fishery was uncertain (these were allocated to the purse seine fisheries and resulted in an increased number of effective releases), and e) the CPUE indices were based on the results of a spatio-temporal analysis using VAST.
3. The change to the estimated relative spawning potential is due to changes to both the spawning potential that occurred due to the fishery and the unfished spawning potential, with some changes to the assessment affecting spawning potential more and others unfished spawning potential more.
4. It is not possible to fully understand why the 2020 assessment is markedly more optimistic than the 2017 assessment because the bridging analyses changed more than a single aspect of the model at a time and the effect of the changes depend on the current model configuration. Analyses based on the 2020 diagnostic model suggest that the main causes for the more optimistic results are the use of the conditional age-at-length data (i.e., from the new otolith age and length dataset), the change to the way tag recaptures were classified as being within the assumed mixing period, assumptions around fishery groupings for selectivity, and to a lesser degree addition of new data collected since 2017. Other changes such as natural mortality and fecundity also had sizable effects, but led to more pessimistic results.
5. The change to the method used to determine tagger related mortality led to a 20% increase in releases with only a corresponding 2% increase in recaptures. This had only a small positive effect on depletion and current spawning potential, but had a large impact on estimates of early biomass.
6. The model provides poor fits to the length-composition data, which might lead to it removing the incorrect number of animals from the population, particularly for fisheries with data in weight. The Panel provides suggestions to rectify this, but some of these will involve changes to MULTIFAN-CL.
7. The treatment of the tagging data between the 2017 and 2020 assessments appears to have been a major contributor to the change in the perception of status (and the improved condition relative to unfished). The impact of each change to the tagging data set needs to be better understood than was possible during the review meeting. Moreover, it is important to better understand the nature of the tagging data as well as how it needs to be modelled (e.g., how to address tag mixing, tag-reporting and tagging mortality) – this will be a substantial exercise.
8. There appears to be conflicting signals on absolute abundance based on the data (index, composition, tagging and conditional age-at-length) and how they have been used to fit the model.
9. The analysts should prepare for a transition to an alternative assessment platform given that recommended enhancements to MULTIFAN-CL are unlikely to occur given resources and the retirement of the lead developer.
10. The next assessment should start with the construction of a conceptual model of the system based on qualitative and quantitative information, and apply analytical methods to inform the definition of regions and fisheries.
11. The factorial grid approach to summarizing uncertainty remains state-of-the-art but there are several ways in which the implementation for the 2020 assessment could be refined for the 2023 and future

assessments.

12. The Panel identified several areas where collection of additional data will be beneficial, as well as suggestions for methodological improvements and further data analyses.

Introduction and General Issues

The Panel (see Appendix A for panel biographies) conducted a review of the 2020 assessment of yellowfin tuna (YFT) for the western and central Pacific, including the data inputs, the settings for the diagnostic model and the settings defining the uncertainty grid based on the final Terms of Reference (Appendix B). Prior to the meeting Dr Arni Magnusson provided an annotated set of key questions related to a variety of topics related to each ToR (Appendix C). The Panel was provided with a set of background documents (Appendix D) prior to the meeting of the Panel, as well as a Rshiny-based viewer for output.

The review meeting took place between 7 and 13 September 2022 at SPC, Noumea, New Caledonia, and was chaired by Dr André Punt. The analysts (see Appendix E) gave a presentation of the 2020 assessment and responded to questions from the Panel on the first day of the review. The Panel identified several requests for additional model runs and data analyses that the analysts addressed between meeting sessions. During the subsequent days, the Panel evaluated the responses to its requests (Appendix F), and reviewed the background documents. The conclusions and recommendations from the draft report were presented to the analysts on 13 September 2022, and the meeting report was finalized after the review meeting.

The Panel focused on the model inputs, particularly as they relate to the trend in spawning potential, to the trend in dynamic unfished spawning potential and the ratio of these two quantities. It also examined the assumptions made in model construction and quantifying uncertainty. The diagnostics used to evaluate model fit were examined and additional diagnostics proposed, along with refinements to existing diagnostics. Finally, the review examined the proposed changes to MULTIFAN-CL and identified a range of prioritized research recommendations.

The Panel spent considerable time with the analysts to explore why the 2020 assessment is more optimistic than the 2017 assessment, with a focus on the changes among model runs within the bridging analysis. The previous assessment involved considerable work and many analyses. However, the changes made while conducting the bridging analyses often included changes other than those mentioned in the documentation. For example, two major (or perceived to be major) changes were expected to have been made between models Age10LW (changing the values of the length-weight parameters and the plus-group age) and CondAge, which led to more optimistic[4] results. However, investigation of the input files revealed that *M*-at-age and reproductive-outputs-at-age were also changed when model Age10LW was modified. Consequently, the Panel focused on how modifications to the 2020 diagnostic model would affect the model results (generally the time-series of recruitment, spawning potential, dynamic unfished spawning potential and relative spawning potential). Core conclusions from this exercise are (see Figure 1 for a plot of the results of the 2020 diagnostic model and changes to that model based on changing one aspect of its specification at a time):

- The effects of changing the plus-group age and the length-weight regression parameters are largely inconsequential.
- The change from the 2017 *M*-at-age vector to the 2020 *M*-at-age vector led to more pessimistic results.
- The change from the 2017 reproductive output-at-age vector to the 2020 reproductive output-at-age vector led to more pessimistic results.
- Including the conditional age-at-length data (along with not estimating the deviations in length-at-age for the first 8 quarters) led to more optimistic results.
- The change from the 2017 to the 2020 tagging data set led to more optimistic results. While several changes, including the addition of new data, were made to the tag-recapture data, it was found that using the 2017 tagging data in the 2020 diagnostic model led to a more pessimistic depletion outcome, and using the 2020 tagging data, but restricting it to the tag release groups available for the 2017 assessment produced a more optimistic depletion outcome. This implies that results from the treatment of the tagging data (mixing period allocation and the change to tagger-related mortality) rather than from the addition of new tagging data in 2020 were most consequential. However, additional model runs are needed to determine relative impact of these two changes.

[4] In the sense that relative spawning potential is higher.

The review process was challenged by limited specifications of incremental steps on what changed between the 2017 and 2020 assessments. This appears to be due to constraints on the time available to conduct the assessment and to provide adequate details in the document. Many model and data aspects changed, which meant several bridging steps were combined to save time (individual model runs apparently took over 30 hours). The Panel consequently **recommends** that the analysts be given more time or additional technical support to ensure that model exploration is such that it is possible to fully understand the causes of changes in model results. A step in a bridging analysis should involve a single change only, with perhaps "minor" changes to the MULTIFAN-CL settings that are needed to ensure convergence. Additionally, these changes need to be clearly documented. Table 1 provides a simplified example format that may aid in documenting changes to model specifications. The Panel **recommends** a table like this (or something similar) be adopted for future assessments so that effects can easily be understood and isolated.

The Panel wishes to thank the SPC for hosting the meeting, the thorough background information provided prior to the review meeting, and the participants for the excellent and constructive atmosphere during the review meeting. The Panel particularly wishes to thank the analysts for their skill in addressing the many requests from the Panel quickly and for their considerable patience. The availability of results and analyses during the review meeting substantially enhanced the Panel's ability to address its ToR. In conclusion, the Panel recognizes that the current assessment, while it can be improved, is state of the art. The Panel was impressed by the comprehensive analyses that allowed it to explore a variety of model specifications and test several assumptions.

Panel Deliberations Relative to Each TOR

A. The adequacy and appropriateness of data sources and data inputs to the stock assessment

A.1 Natural mortality

Natural mortality (*M*) is an important quantity that determines spawning potential and its trends. The Panel noted that the pattern of age-specific *M* differs substantially between the 2017 and 2020 diagnostic models. The pattern of *M* with age for the 2017 and 2020 diagnostic models (as well as all yellowfin and bigeye assessments since 2009) used the same approach, as described in Hoyle et al. (2009). However, some difference in the input data for calculating *M*-at-age resulted in different patterns. The growth curve and sex-ratio-at-length vector are key inputs for calculating *M*-at-age and both these inputs changed between the 2017 and 2020 assessments. The differences in *M*-at-age implied that there are essentially no females aged[5] 28 quarters and older in the 2017 model, whereas for the 2020 diagnostic model there continues to be females until the maximum age of 40 quarters based on the new otolith age data (Fig. 9 of Vincent et al., 2020a). The Panel endorsed the general approach for estimating *M*-at-age. This involves fitting a model that depends on empirical data on the sex-ratio at length, a growth curve, a base *M* for males, and assumptions on critical lengths and a multiplier that determines the linear decline in *M* for young ages to the base *M*, plus length at which female mortality begins to increase (Hoyle et al., 2009; Vincent et al., 2020b). The *M*-at-age vector is estimated outside of the stock assessment even though its calculation depends on growth. Thus, *M*-at-age is sensitive to the growth curve, which can ultimately impact biomass scaling. However, until MULTIFAN-CL can be extended to calculate *M*-at-age from sex ratio-at-length internally, the current approach is the most appropriate notwithstanding that a male-biased sex-ratio in larger fish (assumed due to higher female mortality due to reproductive stress) could be explained by dimorphic growth, selectivity or biased sampling (the Panel considered the evidence in favor of differential *M* by sex being the most plausible explanation based on weight of evidence).

The Panel **recommends** continuing the current approach with the base *M* for the Hoyle et al. (2009) method set to 0.2 quarter^{-1} but including alternative values for base *M* in the uncertainty grid. The range of base *M* values could be determined using a likelihood profile or the bounds from Hoyle et al. (2023), but for now there is no basis to set this value other than to the default of 0.2 quarter^{-1}. The *M*-at-age vector in the 2020 assessment had an average *M* of 0.23 quarter^{-1}. Attempts to estimate the average *M* within the model (Request X) led to the *M*-at-age vector having an average *M* of 0.2 quarter^{-1}, with a base *M* of 0.17 quarter^{-1}, but the Panel noted that the many model conflicts meant that this estimate was likely not very robust.

[5] For simplicity ages in this report relate to quarters so an animal of "age 28" is actually seven years old.

A.2 Maturity

Reproductive output-at-age was determined for the 2020 assessment by first computing reproductive output at length from maturity-at-length, fecundity-at-length, and the sex-ratio-at-length (Fig 8 of Vincent et al., 2020b). In the previous assessment (Tremblay Boyer et al., 2017) the proportion spawning at length (spawning fraction) was also used, but as this was from old data collected from the Eastern Pacific. It was decided that these data were not representative of the WCPO model region and were not used in the 2020 assessment. Unlike the 2017 assessment, where the reproductive output-at-length was converted to -at-age externally, the newer version of MFCL used in the 2020 assessment made this conversion within the model using the specified growth curve (external or internally estimated growth from conditional age-at-length data). Assumptions about maturity do not affect the process of fitting the model (the stock-recruitment relationship is included with a weak penalty so it does not influence the results) but directly determine spawning potential, which is used to calculate reference point values. The Panel notes that the 2020 diagnostic model was based on an updated vector of sex-ratio-at-length for the WCPO and differed substantially from that used in the 2017 assessment. The Panel notes that future assessments of yellowfin tuna and ideally other species should provide consistent output on how changes in assumptions (e.g., natural mortality) affect sex ratios and reproductive output (and vice versa).

A.3 Growth

Growth (length-at-age and its variation) is integral to the assessment as two of the primary data sources are length-frequency and weight-frequency data. Because calculation of both *M*-at-age and reproductive output-at-age are based on estimates at length converted to age, they also depend strongly on length-at-age. The 2020 diagnostic model involved estimating a von Bertalanffy growth curve (with no offsets) and including the conditional age-at-length data when fitting the model. The uncertainty grid also included models in which growth was governed by a Richards curve (fitted outside the model) and in which length-at-age was governed by von Bertalanffy curve with offsets for young ages (1-8 quarters). The latter model was estimated from weight- and length-composition data by initially running a model with very high weight on the composition data, then using the resulting estimated growth model as an external input to the assessment model, which then downweighted the composition data.

In relation to the conditional age-at-length data, it was noted that the fitting procedure did not consider age-reading error (not possible within the current version of MULTIFAN-CL) and was assigned high weight (each otolith was treated as an independent observation). The sampling process for age data aimed to obtain a similar number of otoliths per length-class, which means that a growth curve fitted to these data outside the assessment will be biased. The Panel therefore **recommends** not basing the growth curve on an external estimate unless internal estimates are clearly implausible or an appropriate sampling approach to obtain representative population length-at-age data can be developed and implemented, or a growth curve is externally estimated using conditional age-at-length data. Moreover, some otoliths are discarded owing to difficulties with reading – care needs to be taken that this does not bias the distribution of ages within length-classes. The Panel also outlined an alternative approach to show the residuals of the fits to the conditional age-at-length data and explore if there is spatial variation in growth.

There is evidence from both the otolith data and modes in the length- and weight-composition data that growth varies spatially. In particular, it appears from residual patterns of the fits to the conditional age-at-length data that growth is slower in the north and south regions compared to the equatorial regions. The modes in the composition data also indicate slower growth at younger ages, which may be linked to the estimates of the offset parameters for young ages in the growth curves estimated from length composition modes. Conflict was found between the growth increment data from tagging and the otolith data in the external analysis and between the otolith data and the modes in the length/weight composition data in the assessment. Further investigation and data collection is needed to determine the cause of these differences (e.g., spatial, selectivity, seasonal growth).

A.4 Size composition

The current weighting scheme gives equal weight to all composition data sets with a sample size (number of fish measured) of at least 1,000. Moreover, the weighting scheme does not account for how the samples were collected (e.g., large numbers from a few sets/trips or small numbers from many sets/trips). Consideration should be given to weighting the composition data using a metric that reflects the likely

information content of the data (such as sets/trips), but this will require access to more basic data than is currently available to the analysts.

The sample size used in the likelihood for composition data can have a substantial effect on the results of the stock assessment. Therefore, it should be considered carefully. There are various components to including composition data in a likelihood such as the initial sample size input into the stock assessment, the relative sample size among years for a particular fishery, and the method used to scale the sample size with respect to the fit to the data. Fish caught by a particular gear can have substantial correlation in their age or size in a set, trip, or a time-spatial strata. This pseudo replication reduces the effective sample size of the composition data, may differ among fisheries, and can really only be evaluated outside the assessment before the composition data are assembled for use in the assessment, except in a very broad sense. Therefore, the sample size of the composition data should be analyzed outside the stock assessment model (e.g., using bootstrap analysis or spatio-temporal models) or the appropriate measure of sample size chosen (e.g., number of sets or trips). The IATTC uses the number of purse seine sets as the input sample size in their stock assessments as most sets are on schools of similar sized tuna. The effective sample size can then be estimated based on the fit to the model using iterative approaches (McAllister and Ianelli, 1997; Francis, 2011) or approaches that estimate the effective sample size as a parameter. These approaches generally assume that the relative among-year sample size is maintained, but modifications to the maximum have been proposed (e.g., asymptotic functions). The best approach has yet to be determined and is still an active topic of research.

A.5 CPUE Data

The Panel discussed several issues related to the use of CPUE data in the assessment. In particular, it was noted that the model does not fit the mean weights for the index fisheries for regions 3, 4 and particularly 8 well. The reasons for these discrepancies are unclear but may be related to the assumption that selectivity and catchability are assumed to be the same for all regions. Some relaxation of this may be necessary to resolve this (see Section B.1 below).

The CPUE data used in the assessment are based on the use of the VAST framework (Ducharme-Barth and Vincent, 2020). This framework has the advantage that it can provide indices of abundance for the entire area covered in the assessment, and account for spatial correlation, which could be relevant for spatial grids with few data. The Panel supports the continued use of spatial distribution models (recognizing that they were not designed to address issues of preferential sampling within the spatial grids) and provides suggestions to refine the spatial distribution modelling (perhaps using the sdmTMB package as well as VAST).

The choice of cells to use in the CPUE analysis is still a topic of research. The main concern is cells on the edge of the fishery that are fished in some years or quarters and not in others. It is unknown if these cells have low CPUE and that is why they are not fished or if they are not fished for reasons unrelated to CPUE and therefore have CPUE like neighboring cells. One approach is to use cells that are fished for a majority of the modelling period. However, this ignores spatial changes in the distribution due to infrequent environmental events. Xu and Lennert-Cody (2022) restricted the spatial domain of the catch and effort data to the "core" fishing ground for skipjack, which was defined for the OBJ and NOA fisheries as all 1° x 1° squares in the eastern Pacific Ocean with at least 11 and 6 years of CPUE data between 2000-2021, respectively.

Effort creep, to some degree, is likely to have occurred for the fisheries in this assessment, particularly given the developments of gear, vessels and fish finding technology over the period being considered. However, the rates and temporal/fleet specific dynamics of effort creep for different fisheries (i.e., purse seine and longline) have not been adequately studied to provide recommendations of effort creep scenarios to be modelled. Understanding effort creep for the longline fisheries is most important for the yellowfin tuna assessment as they provide the abundance indices. The Panel suggests that sensitivity analyses should be conducted to explore what levels of effort creep are required to impact management quantities, but that the effort creep scenarios applied in models used for management advice should have a sound basis. Other factors such as targeting or fishing costs can also impact catchability and may not relate to technology improvements. The Panel therefore **recommends** that the SPC is supported to conduct further investigation of effort creep in the longline and purse seine fisheries. This will require support from Distant Water Fishing Nations for catch and effort data provision and information on how operations, vessel features, gear and

technology uptake has changed over time for their fleets. The Panel understands a proposal will be submitted from SC18 to the WCPFC19 for support to study effort creep. This study is based on recommendations from the 2022 WCPO skipjack assessment and would focus on pole and line and purse seine fisheries. The Panel **recommends** that this project be expanded with additional funding to also consider longline fisheries.

The seasonal variability in the indices may not have been captured well (Figures 2 and 3). It is not clear if seasonal catchability is needed or whether seasonal movement needs to be captured better. A seasonal-spatial interaction term should be added to the spatio-temporal model of the CPUE data as a first step for dealing with this issue. The seasonality influence on unsampled or under sampled cells in the spatio-temporal model should be investigated and the spatial coverage of the analysis modified to ensure that areas-seasons where yellowfin are unlikely to be present are not included in the analysis. Consideration of using cluster analysis, hooks per set, vessel ID, and hooks between floats as targeting indicators should be continued.

Model runs (requests PP and QQ) revealed that the spawning potential and spawning potential depletion are robust to changing the scale of the CPUE indices (although this did lead to changes in the relative spatial distribution of biomass). Furthermore, downweighting the composition data did not affect spawning potential and its depletion. This suggests that the tagging data are informative with respect to overall population size, but this needs to be explored further for the 2023 assessment.

A.6 Tagging Data

Tagging data are integral to the assessment. The 2017 assessment included fewer tags than the 2020 assessment and identified tags that may not have fully mixed into the population differently than the 2020 assessment. The method used to calculate tagger-related mortality (tagger effects) was changed between the 2017 and 2020 assessments, and this increased the number of releases included in the 2020 models. Another concern relates to the tag-reporting rates. Some of these are informed by tag seeding experiments (although the estimates from the assessment may differ substantially from the prior values) but others are assigned non-informative priors. The Panel explored some alternative specifications for the tag-reporting rates (see Requests AA and BB). There is a need for more tag-seeding experiments, which should help with estimates for several regions/fisheries that were on their upper bounds. These efforts should be prioritized according to where catches are most important. Fisheries with multiple tag-reporting rates over time could be treated as multiple fisheries, each with a time-invariant tag-reporting rate or MULTIFAN-CL could be modified to allow for time-variation in the tag-reporting rates.

The Panel agrees that the approach of using the actual time at liberty for classifying recaptures as mixed or not mixed is more reasonable than simply considering their release and recapture quarters. This changes the number of tags included in the assessment that influence the estimation of fishing mortality. Depending on the assumed mixing period, which was fixed at two quarters in the 2020 diagnostic model, tag recaptures are excluded from the estimation of fishing mortality - the longer the mixing period the more tags are excluded. It appears that changing from the 2017 treatment (that considered release and recapture quarter and not actual time at liberty) to using 182 days and enforcing this by using actual times at liberty resulted in a more optimistic outcome in terms of the dynamic spawning potential depletion, but this needs to checked further as part of the 2023 assessment. However, the basis for using 1-, 2- or even 3-quarters for a mixing period needs more support as the time for tagged animals to fully mix into the population likely varies by season and area. Moreover, results in the assessment report and analyses conducted during the workshop illustrate a conflict between the tagging data (which suggest a more depleted stock) and some of the other data (see Appendix A of Vincent et al., 2020a). Thus, it is necessary to better understand the most appropriate way to set the time before tags are fully mixed into the population that also considers the likely spatio-temporal variation in mixing among tag release events and fish size. Work based on individual-based modelling (IBM) for skipjack tuna by Scutt Phillip et al. (2022) provides a more defensible basis for assigning mixing periods, rather than relying on fixed assumptions for all tag releases. However, the IBM models require a model of movement across age groups (for skipjack based on SEAPODYM) and such a model would need to be developed for yellowfin.

Additionally, the impact of adding the Japanese tagging data set coincided with modifications to the approach for assigning tag mixing. Because of this concern (and the fact that the stock status changed substantively with this model modification) the analysts separated out the impact of the component parts and found that the new method of dealing with the mixing period assignment affected the status the most.

It is also noted that the application of the new model for estimating tagger effects on release mortality occurred very early in the bridging analysis (*Update* step in Figs. 13 and 14 of Vincent et al. 2020a) and had a minor effect on the stock status estimates. However, other changes were also made such as including recaptures that occurred within the mixing period that had capture location and date information but lacked information on recapture fishery. These recaptures were allocated to the purse seine fishery in the release region, but to account for these recaptures the effective releases were required to be increased. The change to the method to determine tagger-related mortality led to a 20% increase in releases with only a corresponding 2% increase in recaptures. This had only a small positive effect on depletion and current spawning potential, but had a large impact of early biomass estimates. It was not possible to fully explore which aspects of the changes to the tagging data has the largest impact on the results and this should be explored during the development of the 2023 assessment.

A.6 Catches

There is uncertainty in the Philippines and Indonesia catches prior to 1990 and investigation of approaches to improve these estimates or include the uncertainty in the assessment should be continued, perhaps through the WPEA (West Pacific East Asia) project. Other tRFMOs receive longline catch in weight or numbers in different years and it should be confirmed that the data received by SPC is received in the units that catches were measured in and not pre-converted by the member states.

B. Model configuration, assumptions, and settings

B.1 Selectivity parameterization

The 2020 diagnostic model has poor fits to most of the length- and weight-composition data when aggregated across all years or when observed and model-predicted mean length or weight are compared (e.g., Figure 18 of Vincent et al., 2020a). Individual year/quarter fits are even worse. The effective sample size used is low (maximum of the numbers sampled are 1,000 divided by 60), but the use of flexible spline selectivities should still provide better fits to the data. The 2017 diagnostic model led to better fits to the composition data used in that assessment, although the composition data were assigned greater weight in that model. A few fisheries with limited composition data share selectivities with similar fisheries, but the misfit was not limited to these fisheries. For some fisheries the aggregated composition data have multiple modes, shoulders, and means that change over time. These all indicate that multiple "fisheries" are being combined or that the fishery selectivity is changing over time, that the fish may not be removed at the correct size, and that information on abundance from the composition data may be compromised. It is important to specify selectivity at least approximately right to remove fish at the correct size, particularly for fisheries that lead to a majority of the removals and to represent the correct sizes for indices of abundance. A more structured approach is needed to model the composition data and selectivity. The Panel **recommends** the following approach:

- Define fisheries using a regression tree analysis applied to the composition data (e.g., Lennert-Cody et al. 2010, 2013; Maunder et al., 2022).
- Describe the fisheries, including the magnitude of catch, sample size, and whether it is an index.
- Triage the composition data to remove data that are likely to be unrepresentative and/or unreliable. This may include excluding data for a whole fishery (and sharing selectivity), entire years, or for some lengths (e.g., when small amounts of fish under the minimum legal size are caught).
- Avoid aggregated compositions that show multiple modes, shoulders, or other unusual patterns (i.e., not logistic or double normal) by separating them into more fisheries or allowing for time-varying selectivity.
- Assume that selectivity is length-based unless it is known to be age-based (e.g., due to ontogenetic movement).
- Ensure that the composition data for fisheries that catch a large proportion of the catch are fit well. This might require the use of flexible time-varying selectivity.
- Consider downweighting the composition data for fisheries with low catch as these do not need to be fit well.
- Fit the composition data for indices well - consideration should be given to allowing selectivity to be more flexible and time-varying if necessary.

- Use the empirical selectivity diagnostic (Maunder et al., 2020)[6] to check that selectivities are appropriate.
- Use the empirical selectivity method to determine the number of knots and their position when using splines.
- Consider dropping the composition data for fisheries whose selectivities are shared with those for another fishery because the data for those fisheries are considered inadequate, particularly if it has low catch.

The index fisheries share catchability and selectivity. This allows the CPUE data to provide information on scaling among the regions. Sharing selectivity among regions may be problematic because there is evidence of differences in length-at-age among regions, which might be best modelled using different selectivity patterns for each region. However, different selectivity patterns also imply different catchability, which, if modelled, will lose the information on relative regional scaling. Consideration should be given to allowing for some differences among regions while maintaining similarities to retain information on regional scaling. For example, catchability and selectivity for each region could be modelled as a penalized deviate from the overall mean or one region set as the reference catchability and selectivity and the other regions deviating from that region. Selectivity might require age-/length-specific deviates or something more complicated with either the peak changing or a functional form describing offsets for all ages/sizes. It is unknown if this is possible in MULTIFAN-CL.

B.2 Recruitment

The assessment model fixes the recruitment for the recent six quarters to the mean. This may influence the results if there is information about recruitment in the data. The model should be run estimating these recruitments to determine the impact on the results. When using the penalized likelihood approach, a log-normal bias-correction factor is needed to ensure the deterministic equation represents the expected value (mean) of recruitment. However, when information is limited for a particular year, the full bias correction will bias the estimates. In the extreme case of no information, the bias will be equal to the bias-correction factor $-\sigma_R^2/2$) and the bias-correction factor should not be used. Lack of information can occur in early years due to the lack of composition data and in recent years because some cohorts are included in the composition data for only a few years. Since there is no data in the projections, the bias-correction factor should not be used for future recruitments. For years with partial information only a partial bias correction should be used and this is described by the bias-correction ramp in Stock Synthesis (Methot and Taylor, 2011) and should be determined for both the left hand (early years) and right hand (recent years) sides.

Approaches that treat recruitment as a random variable (e.g., random effects, state-space, Bayesian) do not require the bias correction ramp. However, these approaches require integration and are often not practical for complex stock assessments or are not available in the software used.

As previously mentioned, there may not be information on the recruitments for the early years because the model starts after some of the associated cohorts are part of the catch. It is tempting to set the recruitment deviates for these years to zero (in combination with removing the log-normal bias-correction factor) and making the recruitment equal to that expected from the stock-recruitment relationship (or the average). However, this will ignore any uncertainty in the recruitment and prevent the model from estimating any long-term trends in recruitment. On the other hand, if the recruitment deviates are estimated, they can compensate for a model misspecification. This approach is also associated with the selection of the start time of the model and the method used to create the initial age-structure. The best approach has yet to be determined generically, and requires further research.

B.3 Movement

There is a possibility that movement differs between adults and juveniles. Future work should look at releases and recaptures by size groups to identify any differences. Age-specific movement should be investigated in the assessment either by fixing movement to zero for adults or estimating age-specific

[6] Implemented in the R package empirical.selectivity. remotes::install_github("roliveros-ramos/fks") remotes::install_github("roliveros-ramos/empirical.selectivity")

movement. The assessment report notes that an alternative hypothesis is that spatial differences in growth are currently mis-specified, and the models are attempting to compensate for this through the resulting movement and abundance patterns. Alternative spatial modelling should be considered as described below.

B.4 Hessian matrices

For the models presented during the meeting, the computation of the Hessian was missing, along with the analogous approximate asymptotic variance (and covariance) estimates. Several suggestions were provided on how this might be improved (e.g., modifying the configurations so that parameters were not on the bounds, trying a generalized inverse for the Hessian to obtain correlation estimates). The uncertainty of the parameters of interest over alternatives could perhaps then be used to judge how different models might be combined and/or weighted.

B.5 Model complexity

Model complexity is one of the main reasons the WCPFC and SPC requested an external review. The extent of data and model complexity requires more time and attention than can be afforded via WCPFC's SC process. The model complexity and ability to transparently demonstrate the interplay of how data and model configurations impact results is a theme of this review. This occurred partly because the lead author was unavailable to respond to queries on some of the assessment details (no longer working for SPC, but did provide answers to some questions by e-mail). The SPC analysts have done an admirable job in developing ways to disentangle the interactions between new data and model configurations. During the review, the Panel suggested developing a more transparent way to document incremental model changes (Table 1).

The Panel notes that the current model formulation (e.g., spatial structure and fishery definitions) aims to mimic the structure of the bigeye assessment. This is partly to improve the efficiency of conducting assessments of bigeye and yellowfin side by side and partly related to the similarity in the fisheries targeting the two stocks. However, the structure of the data for, and the behavior of, yellowfin differ from those of bigeye, such that the current model structure for yellowfin likely leads to model instability and unnecessary complexity. The Panel **recommends** the following elements be considered in developing a new (2023) yellowfin assessment:

- Development of a conceptual model for yellowfin in the WCPO. This would involve synthesizing (in the absence of a model) the information on data (e.g., length and CPUE data), what is known about movement from tagging, including whether juveniles would be expected to have the same movement rates as adults, information from genetics and other indicators of stock structure, and information from oceanography on likely distribution. The development of the conceptual model might occur during a workshop or other review process involving relevant experts, including the stock assessment team.
- Analysis of the length-composition data based on methods used by Maunder et al. (2022) – or a similar approach – to assess which areas/fleets should be combined for the purposes of defining regions and fisheries (see Lennert-Cody et al., 2010, 2013).
- In general, the Panel expects that simpler models will result from this process. In particular, there seems little basis for separating region 9 for this assessment. However, it notes that without evaluating a model with the added complexity (e.g., movement among areas), it may be difficult to appreciate the extent to which simpler models may violate assumptions.
- The analysts should use the results of the conceptual model to identify "realism constraints" and expected model behavior. Consideration should be given to in which regions recruitment should be expected (the current result that there is no recruitment to region 8 seems implausible *a priori* – but earlier assessments had zero recruitment in other regions).
- To the extent possible, multiple fisheries based on the same gear within the same region should be avoided, unless needed given difficulties in replicating tag returns.
- The Hessian matrix should be explored to assess not just the variances of the parameter estimates and the derived variables, but also which parameters may be confounded.

Absent the above process, the Panel identifies that there are some model options that would provide worthwhile information:

- Consider a model based on data for only the equatorial areas (regions 7, 8, 3 and 4) modelled as a

single area and compare its results with an equivalently parameterized Stock Synthesis model. The model would be structured as "fleets-as-areas" with the fleets selected using, for example, the regression tree approach outlined above.

- Allow juvenile movement rates to differ from those for adults (which may be set to zero).
- Tagging data are typically complicated by the limited opportunities to tag fish, which restricts the spatial distribution of releases. Consequently, tag mixing is a major problem that needs to be dealt with when analyzing the data. A fine-scale movement model would be useful for defining the time it takes for tagged animals to fully mix into the population within a region and experience the same probability of recapture as untagged fish in the region. A new approach that models the spatial-temporal distribution of tags using advection diffusion models and the spatial distribution of the untagged population using spatio-temporal models is being developed for skipjack tuna in the eastern Pacific Ocean (Maunder et al., 2021; Mildenberger et al., 2022). This fine-scale spatio-temporal approach would avoid the need to eliminate information by estimating a separate fishing mortality parameter during the mixing period and minimize the edge effect related to how close the fish are tagged to the edge of the large blocks used in the assessment model. However, methods to integrate the information from the analysis into the stock assessment model need to be developed.

B.7 Representing uncertainty

The approach to demonstrating assessment uncertainty is primarily done via combinations of assumptions in a "grid" formulation. This approach considers some aspects of structural uncertainty that is otherwise difficult to demonstrate. The Panel noted that the uncertainty grid for the 2020 assessment involved four dimensions (steepness, the growth model, how much the length-composition data are downweighted, and the number of quarters used to define the time until tags are fully mixed into the population). The review identified the value of base natural mortality and the assumptions regarding the regional structure and the number of fisheries as additional key dimensions of uncertainty.

The Panel noted that the construction of an uncertainty grid remains a state-of-the-art way to synthesize uncertainties that cannot be captured in a single run of an assessment model. However, it noted that the selection of the elements of the uncertainty grid needs to reflect the purpose to which the uncertainty analysis will be used (e.g., used to characterize stock status vs as the basis for a management strategy evaluation) and how the weights are assigned to the elements of the grid. The grid (and the *a priori* weights) assigned to the levels of each factor should be determined by the analysts who conducted the assessment. It may be possible to run a subset of all possible combinations of levels depending on the purpose of the analysis and approaches from ensemble modelling, which is studied extensively outside of fisheries science, could be used to summarize the results. Weighting of models within an uncertainty grid can be achieved simply, e.g., by assigning equal weight to each model in the grid, or by assigning 0/1 weights depending on whether the model leads to poor diagnostics or unrealistic outcomes, and or using a broad set of criteria including goodness of fit, balance of the factors, etc.

In other RFMOs structural uncertainty is sometimes evaluated through the use of a "reference grid". For example, as part of the work to condition the operating model for testing management procedures, the CCSBT analysts identified key uncertainties that were then treated as included in the cross of some 432 different model configurations (CCSBT, 2020). For the management procedure testing, some parameters were drawn from prior distributions whereas others were drawn from a quasi-marginal posterior distribution. This provides an approach for evaluating uncertainty where the analysts' assumptions (i.e., the priors) and other parameters are clearly specified.

The IATTC recently implemented a risk assessment to evaluate probability statements imbedded in the harvest control rule. The risk assessment is based on alternative hypotheses about how to overcome issues in the stock assessment. This pragmatic approach (Maunder et al., 2020) is a compromise between computational demands, complexity, and statistical rigor. It acknowledges the need to weight models based on information in the available data but does so in a context where the complexity of fisheries stock assessment models prevents strict adherence to statistical rigor. The main features of this approach are: 1) hypotheses about states of nature are represented by alternative stock assessment models with specific model structure, data use and parameters; 2) hypotheses are grouped into a hierarchical framework, which highlights similarities among models thereby avoiding that any one hypothesis, or overarching hypothesis,

inadvertently dominates the outcome of the risk analysis, and facilitates model development and weight assignment; 3) sub-hypotheses represent models with parameters that cannot be reliably estimated within the assessment model and are therefore fixed in the models; 4) multiple metrics are used to evaluate the reliability of the models and the plausibility of the hypotheses they represent; 5) model fit only plays a limited role in the metrics used to evaluate models; 6) and an efficient approach to eliminate unlikely hypotheses. The approach has been applied to the bigeye and yellowfin stock assessments. The metrics used to weight the models are mainly based on diagnostics, but also include fit to the data and realism of parameter estimates and results. Each metric was evaluated subjectively by a panel of experts (the stock assessment team) and the results averaged.

A more objective, transparent, and automated approach for model weighting is desired. Several workshops have been held, or will be held soon, and are designed to improve the IATTC approach. A virtual workshop on Model Diagnostics in Integrated Stock Assessments was held during Jan 31-Feb 3, 2022. A virtual workshop on Model Weighting will be held on 28 Nov – 2 Dec, 2022. A further workshop on weighting for tuna models will be held in New Zealand during 5-10 March 2023. The final approach will be developed based on the output of these workshops, but early indications suggest that diagnostics will be used to limit the models included in the ensemble and included models will be weighted by the prediction ability of the models. The Panel suggests it will be beneficial for SPC staff to attend these workshops and for final decisions on modelling and summarizing uncertainty deferred until after the workshops.

C. Model diagnostics

The Panel reviewed the various approaches for displaying the results of the assessment, including the Shiny app and the MULTIFAN-CL viewer. Over the course of the workshop, it became apparent that there were a variety of tools that different analysts used and favored for their work. Some favored specialized display and processing code that they wrote primarily for their own purposes. The push by SPC to develop more transparent dashboards to easily diagnose model results for different configurations is an excellent step in that direction.

Specific to model diagnostics, the Panel **commends** the analysts for the breadth of methods available to view results. The Panel had several suggestions regarding diagnostics:

- The approach showing the weight- and length-frequency data could be enhanced by showing the model fit as well as the population length-frequencies.
- The method for summarizing the fit to the conditional age-at-length data should reflect how the data were collected. Specifically, the ages were sampled within length-classes but the plots shown suggest that they were conditioned on ages within the population. The approach for presenting such information (and model fits) shown in the package "r4ss" should be considered as an option (the results of Request W is the step in this direction).
- More diagnostics are needed for the tagging data including:
    - Diagnostics to understand the ability of the model to fit the tagging data (and hence estimate between-region movements) are needed. Such plots should show time-series of observed and model-predicted recaptures (with totals in the legend) by release group (and excluding the recaptures during the mixing period) (e.g., Figure 4).
    - Diagnostics related to the tag mixing. For example, plotting the recapture rates with distance and time from release.
    - Diagnostics to determine what is informing movement, or at least determine if movement is counter to the tagging data. This might involve running the model with and without movement (or cutting the movement in half for all areas) and determining which likelihood functions are impacted.
- Other suggested improvements to the diagnostics are: (a) including a table of the model parameters, indicating which are estimated and which are pre-specified, (b) adding the means over time of the observed and model-predicted mean lengths and weights to the associated plots, and (c) adding plots of model-predicted CPUE vs residuals as well as the time-series plots (e.g., Figures 2 and 3).
- Add time-trajectories of fishing mortality (or exploitation rate) (e.g., Figure 38 of Vincent et al. [2020]) to the Shiny app.
- Report the mean weight for each observed and model-predicted composition by fishery (from the

length-frequency data) (e.g., Figure 5) to assess if the right numbers are being removed given the model fits the catch weight well.
- Add reproductive output curves as a function of both age and length to the Shiny app.
- Clarify how the plot of where biomass by region comes from is calculated. This is potentially a very informative diagnostic, but the current version does not behave as expected.
- Plots of the average spawning potential against the average CPUE (e.g., Figure 6).

D. Recent MULTIFAN-CL model developments

D.1 Catch-conditioned vs catch-error

The current assessment is based on a 'catch-error' formulation in which 'effort deviations' are estimated to ensure a good match to observed catch, with a large penalty placed on deviations between the observed and model-predicted catches. This leads to a large number of effort deviations being estimated. SPC staff have developed an approach based on a Newton-Raphson approach to solving the catch equation, along with a way to specify fishing mortality for quarters*fisheries with missing data, which reduces the number of parameters substantially (there are some additional parameters related to the approach for inferring missing catches). The Panel **endorses** this approach (referred to as catch-conditioned model), which should simplify the models and ideally help to achieve a positive definite Hessian matrix.

D.2 CPUE likelihood

A new likelihood has been developed that allows the estimation of a parameter (the 'overdispersion' parameter) that multiplies the input CV for the CPUE data. The Panel notes that estimating an overdispersion parameter can lead to the model ignoring the CPUE data if there are conflicts in the data and should be used with care. The Panel **recommends** a second option be developed where the variance of logCPUE is the sum of the square of a pre-specified CV and an overdispersion variance.

D.3 The orthogonal-polynomial parameterization of recruitment

Many MULTIFAN-CL models include a large number of recruitment parameters because recruitment is estimated for each combination of year, quarter, and region. The orthogonal-polynomial parameterization of recruitment reduces the number of parameters by specifying log-recruitment as the sum of polynomials in year, quarter, region and quarter*region. Example plots showed that the approach can capture the more complex traditional parameterization. However, the Panel was concerned that this approach added yet another dimension to the model specification process and **recommends** that it only be used for data-poor situations or for the earlier years for assessments of data-rich stocks for which there is often limited information to inform estimates of recruitment.

D.4 The Dirichlet-Multinomial distribution for length- and weight-composition data

The Dirichlet-Multinomial distribution is an extension of the traditional multinomial distribution that permits estimation of an overdispersion parameter (that might reflect the effects of samples that differ from simple random sampling, but also model mis-specification). The Panel endorses this approach but **recommends** that it be considered alongside the robust normal distribution and McAllister-Ianelli tuning. The Panel notes that all methods for weighting composition data depend on 'stage-1' sample sizes and emphasizes the importance of specifying these correctly.

E. Future research areas, in priority order

The Panel identified several research activities and general methodological recommendations which, if addressed, should improve the ability of the assessment to provide scientific advice for management decision making. These recommendations relate to tasks that would require substantial additional work so could not be conducted during the review.

E.0 General

1) There will be benefit for SPC staff to attend the 28 Nov -2 Dec 2022 virtual Model Weighting workshop as well as the workshop on good practices for tuna assessments (that includes weighting for tuna models) and the associated spatial workshop, which will be held in New Zealand in early 2023.
2) Improve diagnostics of all aspects of the model as outlined in section C.

3) Include an operational definition of "convergence" in the assessment. This may relate to jittering outcomes, gradient values, etc.
4) Specify what constitutes an "appreciable" change in model results and use it consistently throughout assessment documents.

E.1 Model inputs

1) The composition data should be weighted using a metric that reflects the likely information content of the data (such as sets/trips), but this will require access to more basic data than is currently available to the analysts. A review of the approach to calculating the initial "stage 1" sample sizes that are input to MULTIFAN-CL will be useful.
2) It should be confirmed that the composition data are being received in the units and sample size that it was recorded in, particularly for the longline fleet. There have also been issues of consistency between data recorded by observers, commercial operations, and training vessels and the source should be thoroughly evaluated before use in the model.
3) There are also some possible issues with the purse seine composition data that should continue to be investigated including the difference between observer and port sampling, observer grab sampling bias correction, and the conversion factor from gilled and gutted to whole weight, which is based on 100 fish.
4) There have been changes in the tag mortality rates due to tagger effects in the yellowfin assessment and the recent skipjack assessment and further investigation is needed to ensure that the best approach is used as this assumption is one of the factors that can have a notable impact on the results.
5) Conduct a review of all assumed tag mortality and tag shedding values so the basis for these values is clear.
6) With regard to constructing a CPUE index:
    a. Run the spatio-temporal model (e.g., VAST) by region, and (a) compare correlation between the regionally estimated indices (independently) with the same regions split up from global model results, (b) compare decorrelation distances among regions and see how different they are from the global estimate, and (c) assess the extent that within-region trends differ from the global trends.
    b. Examine the extent to which the current indices are correlated owing to their being computed from one model and reflect this (if substantial) by a variance-covariance matrix when fitting to the data.
    c. (a) examine if covariates can be categorized by abundance and catchability, (b) determine how covariates affect the model, (c) consider including interaction terms, and (d) include a quarterly random effect, perhaps in a hierarchical approach
    d. Consider running the spatio-temporal model within a (main) region for all fleets and compare the results to those from a run with only a principal fleet included to assess the effects of combining data for multiple fleets into a single analysis.
    e. Further evaluate both the definition of viable cells and how the spatio-temporal model shares information for cell-times with little information. This is particularly important for evaluating the size of the north and south regions and the influence of edge effects in the CPUE standardisation.
7) Investigate approaches to improve the estimates of the catches by the Philippines, Vietnam and Indonesia or include the associated uncertainty in the assessment if the model estimates are sensitive to this uncertainty. It should be confirmed that the data received by SPC are received in the units that they were measured in and not pre-converted by the member states, particularly for the longline fleet.
8) Re-analyze the length-weight data (e.g., by conversion type and season) for use in quantifying the extent of variation in weight-composition due to error about the length-weight relationship.
9) Develop an age-reading error matrix based on double-reads and include this when computing predicted conditional age-at-length data (inclusion of aging error is not an available feature in MULTIFAN-CL). Also explore sensitivity to including all of the age data irrespective of whether the ages are agreed or considered as high confidence

E.2. Model configurations

1) Develop a conceptual model of yellowfin tuna in the WCPO and use this along with the regression tree approaches developed by the IATTC (e.g., Lennert-Cody et al., 2010, 2013; Maunder et al., 2022) to define regions and fisheries.
2) Allow for some differences among regions in selectivity and catchability while maintaining similarities to keep information on regional scaling.
3) Implement a sex-structured version of MULTIFAN-CL. The current version is sex-aggregated, necessitating some complex modelling to outputs such as computing spawning potential.
4) Explore which aspects of the changes to the tagging data had the largest impact on the results of the 2020 assessment by modifying the data used in the 2020 diagnostic model.
5) Examine the releases and recaptures by size groups to identify any differences in movement rate between ages.
6) Check that the observation that effort-based projections in the skipjack assessment impacted estimates of population scale does not occur for the yellowfin assessment.
7) Construct an individual-based model for yellowfin tuna to provide a basis for selecting the time it takes for tags to be fully mixed into the population – this will require the development of a model of movement (e.g., based on an application of SEAPODYM or the approach being developed at IATTC).
8) Explore different starting years for the assessment given the uncertainty regarding past catches.
9) Investigate effort creep in the longline and purse seine fisheries further. This will require support from Distant Water Fishing Nations for catch and effort data provision and information on how operations, vessel features, gear and technology uptake has changed overtime for their fleets.
10) Examine a recapture-conditioned version of the model (e.g., McGarvey and Feenstra, 2002) to allow an exploration of how much the tagging data impact the estimates of fishing mortality and hence spawning potential, and in particular the estimates of the unfished spawning potential.

E.3 Modifications to MULTFAN-CL

1) Implement length-based selectivity. Length-based selectivity seems more natural for the fisheries concerned and the fact that two of the primary data sources are length- and weight-composition data.
2) Extend MULTIFAN-CL so that variability in weight-at-length can be taken into account.
3) Extend MULTIFAN-CL so that it is possible to specify the number of spline knots when defining selectivity and where they are located with respect to age (length) as the current approach means that the selectivity for some knots is constrained to zero.
4) Extend MULTIFAN-CL so that account can be taken of age-reading error when fitting to conditional age-at-length data.
5) Add the ability to specify overdispersion in CPUE as an additive rather than multiplicative factor.
6) Integrate the calculation of *M*-at-age from the sex-ratio data into MULTIFAN-CL unless a sex-specific assessment is used.

E.4 Data collection

1) Access to set-and trip-level data will enhance the ability to weight the length-frequency data because the number of sets/trips will usually be a better measure of the information content of the length- or weight-frequency sample than the number of fish measured.
2) Further investigation and data collection is needed to determine the cause of the differences (e.g., spatial, selectivity, seasonal) between the tagging growth increment data and the otolith data. Conduct age validation studies.
3) Tag-seeding experiments to develop priors for the tag-reporting rates should be conducted for fisheries/regions for which the number of recaptures is high and no previous tag-seeding experiments have been conducted.
4) Continue to collect data on sex-ratio and spawning frequency to enable refinement of the *M*-at-age and reproductive output-at-age vectors.
5) Collection of additional information on the conversion from processed to whole weight is needed to improve the relationship and also allow inclusion of the additional variation in weight-at-age for fitting weight composition data.

6) Enhance the regular collection and aging of otolith data to use as conditional age-at-length data in the stock assessment to improve estimates of growth. These data should be collected broadly across the spatial range of the fishery and size classes. Data to validate the ages should also be collected, e.g., chemical marking of tagged fish.
7) Plan and then start collection of future tissue samples for the application of close-kin mark-recapture methods.

References

CCSBT. 2020. Report of the Eleventh Operating Model and Management Procedure Technical Meeting. https://www.ccsbt.org/sites/default/files/userfiles/file/docs_english/meetings/meeting_reports/ccsbt_27/report_of_OMMP11.pdf

Francis, R.I.C.C. 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. 68,:124–1138.

Hoyle, S.D., Williams, A.J., Minte-Vera, C.V. Maunder, M.N. 2023. Approaches for estimating natural mortality in tuna stock assessments: Application to global yellowfin tuna stocks. Fish. Res. 257, 106498.

Lennert-Cody, C.E., Minami, M., Tomlinson, P.K., Maunder, M.N. 2010. Exploratory analysis of spatial temporal patterns in length frequency data: An example of distributional regression trees. Fish. Res. 102: 323–26. https://doi.org/10.1016/j.fishres.2009.11.014.

Lennert-Cody, C.E., Maunder, M.N., Aires-da-Silva, A., Minami, M. 2013. Defining population spatial units: Simultaneous analysis of frequency distributions and time series. Fish. Res. 139: 85–92. https://doi.org/10.1016/j.fishres.2012.10.001.

Lennert-Cody, C., Valero, J.L., Aires-da-Silva, A., Minte-Vera, C. 2020. Implementing reference point-based fishery harvest control rules within a probabilistic framework that considers multiple hypotheses. DOCUMENT SAC-11 INF-F REV. https ://www.iattc.org/GetAttachment/46edbd8e-22f9-4bb3-8d26-d4cfd24a472c/SAC-11-INF-F_Implementing-risk-analysis.pdf [iattc.org]

McAllister, M.K., Ianelli, J.N. 1997. Bayesian stock assessment using catch-age data and the sampling/importance resampling algorithm. Can. J. Fish. Aquat. Sci. 54: 284–300.

McGarvey, R., Feenstra, J.E. 2002. Estimating rates of fish movement from tag recoveries: conditioning by recapture. Can. J. Fish Aquat. Sci. 59: 1054-1064.

Maunder, M.N., Xu, H., Lennert-Cody, C.E. 2022. Developing fishery definitions for the skipjack tuna stock assessment in the EPS. IATTC Document SAC-13 INF-I.

Maunder, M.N., Xu, H., Lennert-Cody, C.E., Valero, J.L, Aires-da-Silva, A., Minte-Vera, C. 2020. Implementing reference point-based fishery harvest control rules within a probabilistic framework that considers multiple hypotheses. IATTC Document SAC-11 INF-F REV. https://www.iattc.org/GetAttachment/46edbd8e-22f9-4bb3-8d26-d4cfd24a472c/SAC-11-INF-F_Implementing-risk-analysis.pdf

Maunder, M.N., Xu, H., Schaefer, K.M., Fuller, D.W. 2021. Assessment methods for skipjack in the EPO: a proposal relying on recent data from the IATTC regional tuna tagging program (2019-2022). IATTC DOCUMENT SAC-12-06. https ://www.iattc.org/GetAttachment/e9d4f426-edd3-4498-8fc9-e8a647f44f16/SAC-12-06%20-%20Assessment%20methods%20for%20skipjack%20in%20the%20EPO%20using%20tagging%20data [iattc.org]

Methot, R.D., Taylor, I.G., 2011. Adjusting for bias due to variability of estimated recruitments in fishery assessment models. Can. J. Fish. Aquat. Sci. 68: 1744-1760.

Mildenberger, T.K., Nielsen, A., Maunder, M. 2022. Spatiotemporal tagging model for skipjack in the EPO. IATTC DOCUMENT SAC-13-08. https ://www.iattc.org/GetAttachment/a89cea47-8552-4ab7-b6ca-5b4115f2e1c9/SAC-13-08_Spatiotemporal-tagging-model-for-skipjack-in-the-EPO.pdf [iattc.org]

Xu, H., Lennert-Cody, C.E. 2022. Standardizing the purse-seine indices of abundance and associated length compositions for skipjack tuna in the eastern Pacific Ocean. IATTC Document SAC-13 INF-K.

Table 1. Suggested format for a table of model changes. The idea is to simplify ways of showing how different model steps/configurations differ. The idea is to choose an order that makes sense then perhaps test some sequence increments (perhaps reorder a few changes that appeared to have the most difference in model fits or outcomes).

Model name	A	B	C	D	…
Base_2017	X				
Update_2020		X			
Add_comp		X	X		
Add_tag		X	X	X	
…					

Legend:

A) Base 2017 model
B) As A) extend model to 2020 (with only catch totals updated)
C) As B) but with composition data included
D) …

Figures

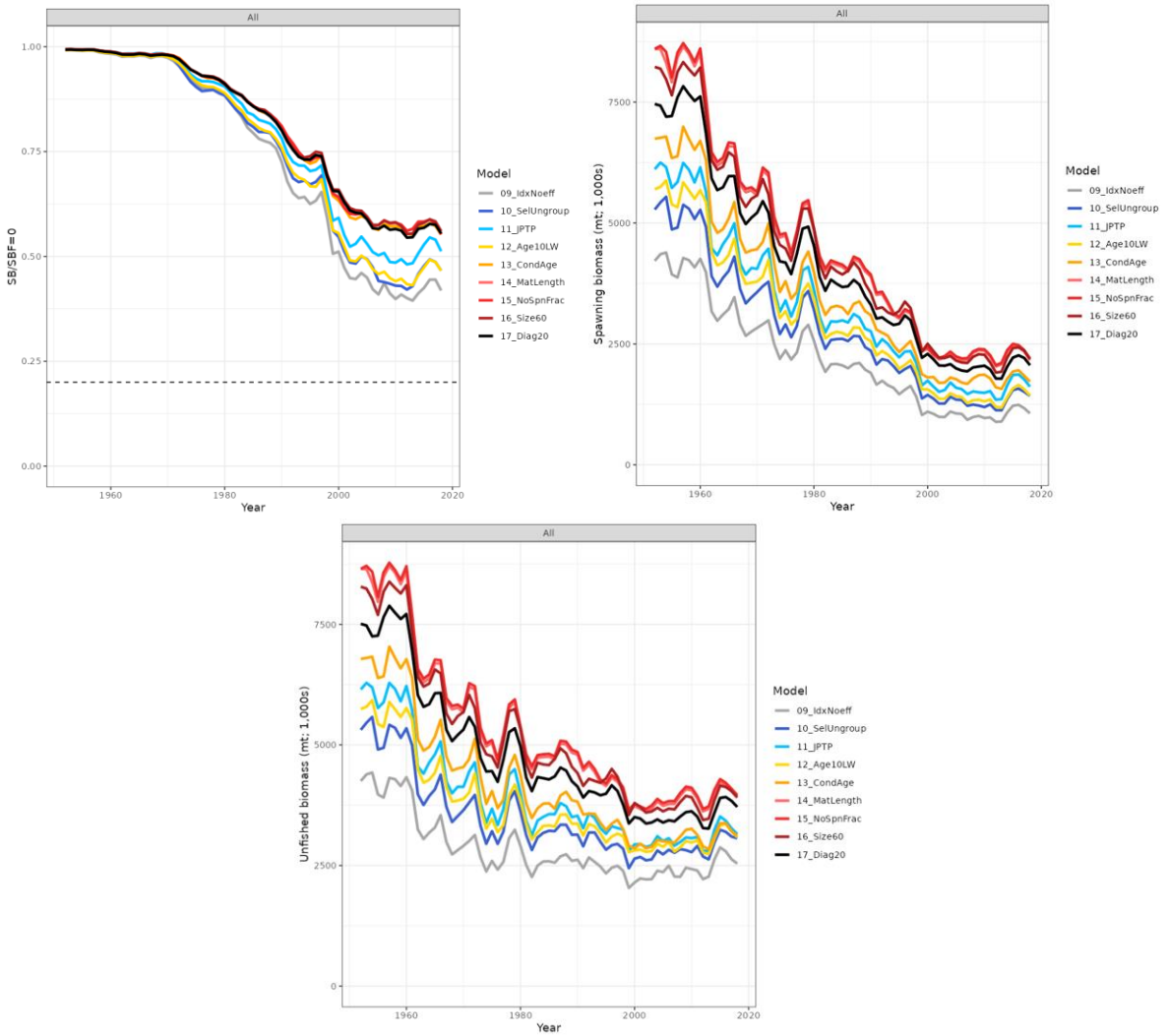


Figure 1. (a) Second phase of the stepwise models where most of the changes from 2017 to 2020 diagnostic models occurred. Descriptions of changes in each step are in Appendix G. 17_Diag20 = 2020 diagnostic model.

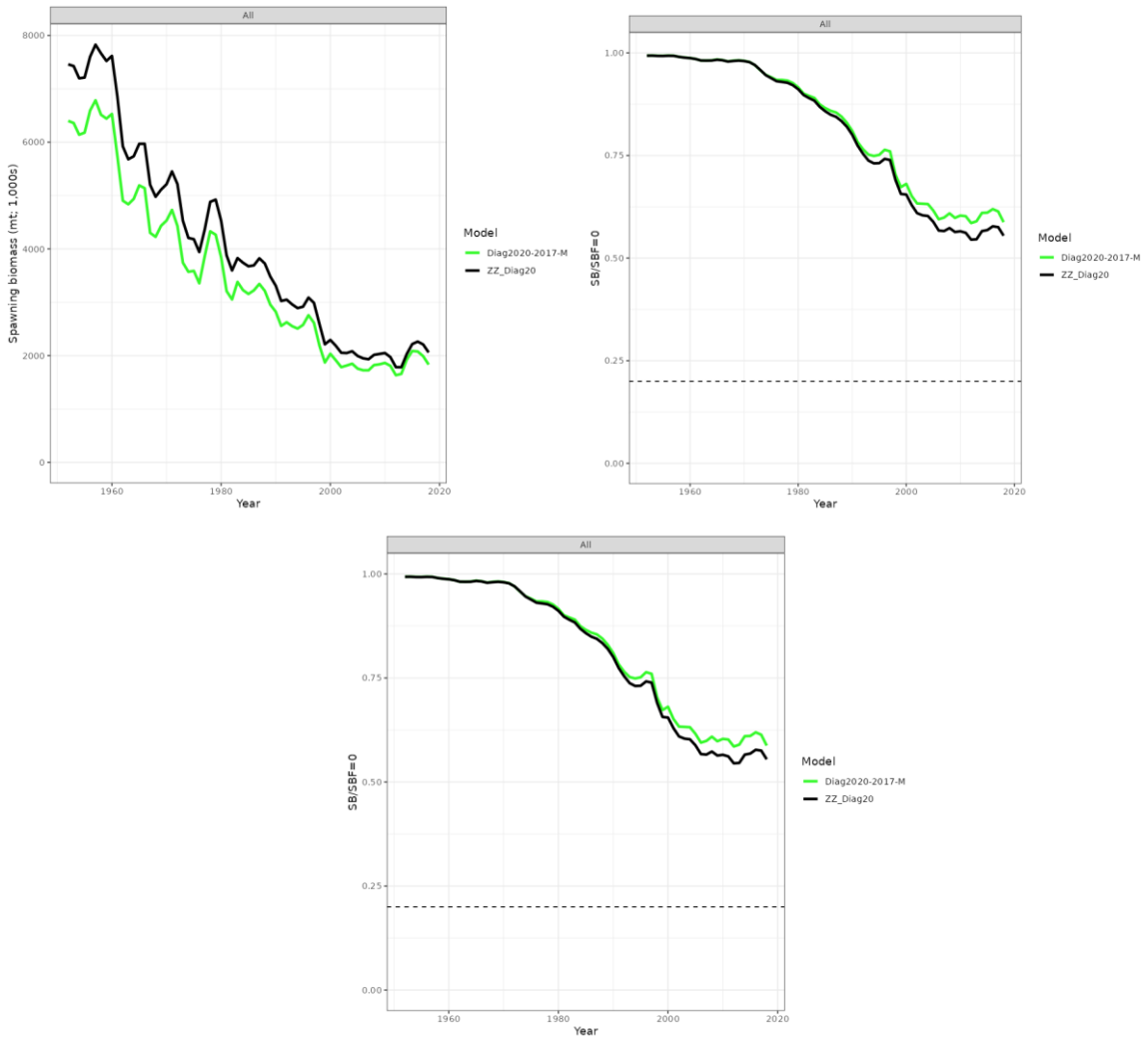


Figure 1(b). Change from the 2017 *M*-at-age vector to the 2020 *M*-at-age vector.

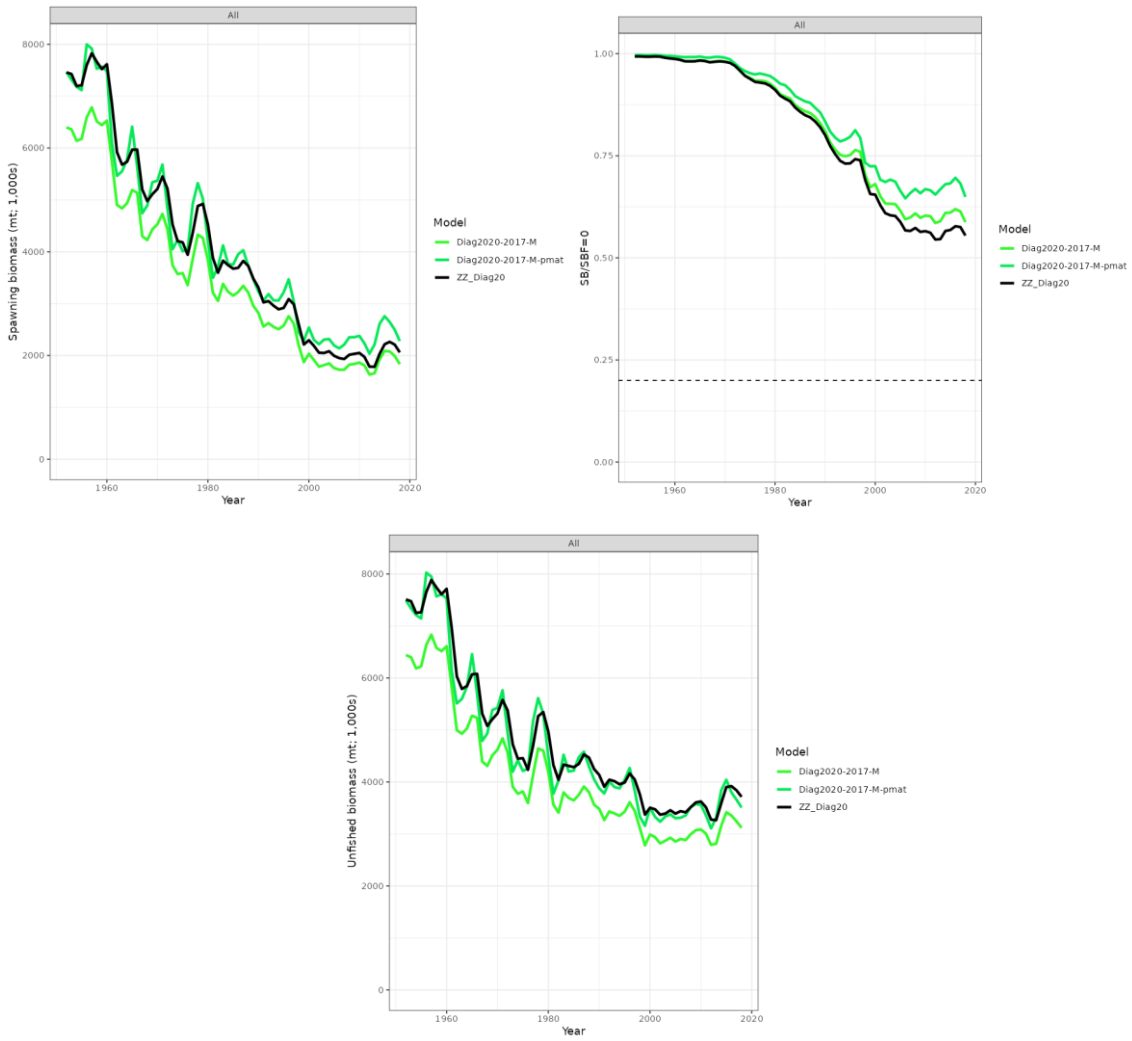


Figure 1(c). Change from the 2017 reproductive output-at-age vector to the 2020 reproductive output-at-age vector.

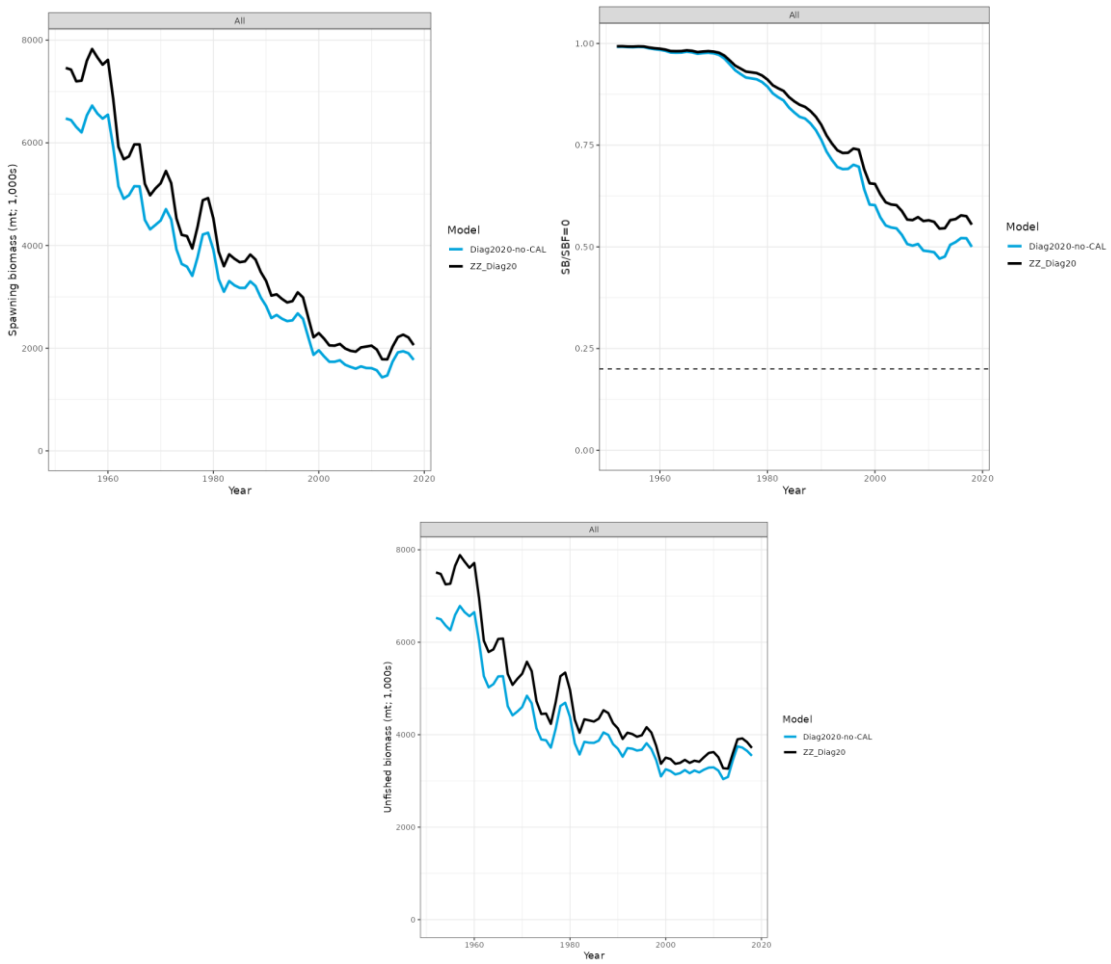


Figure 1(d). Including the conditional age-at-length data (along with not estimating the deviations in length-at-age for the first 8 quarters)

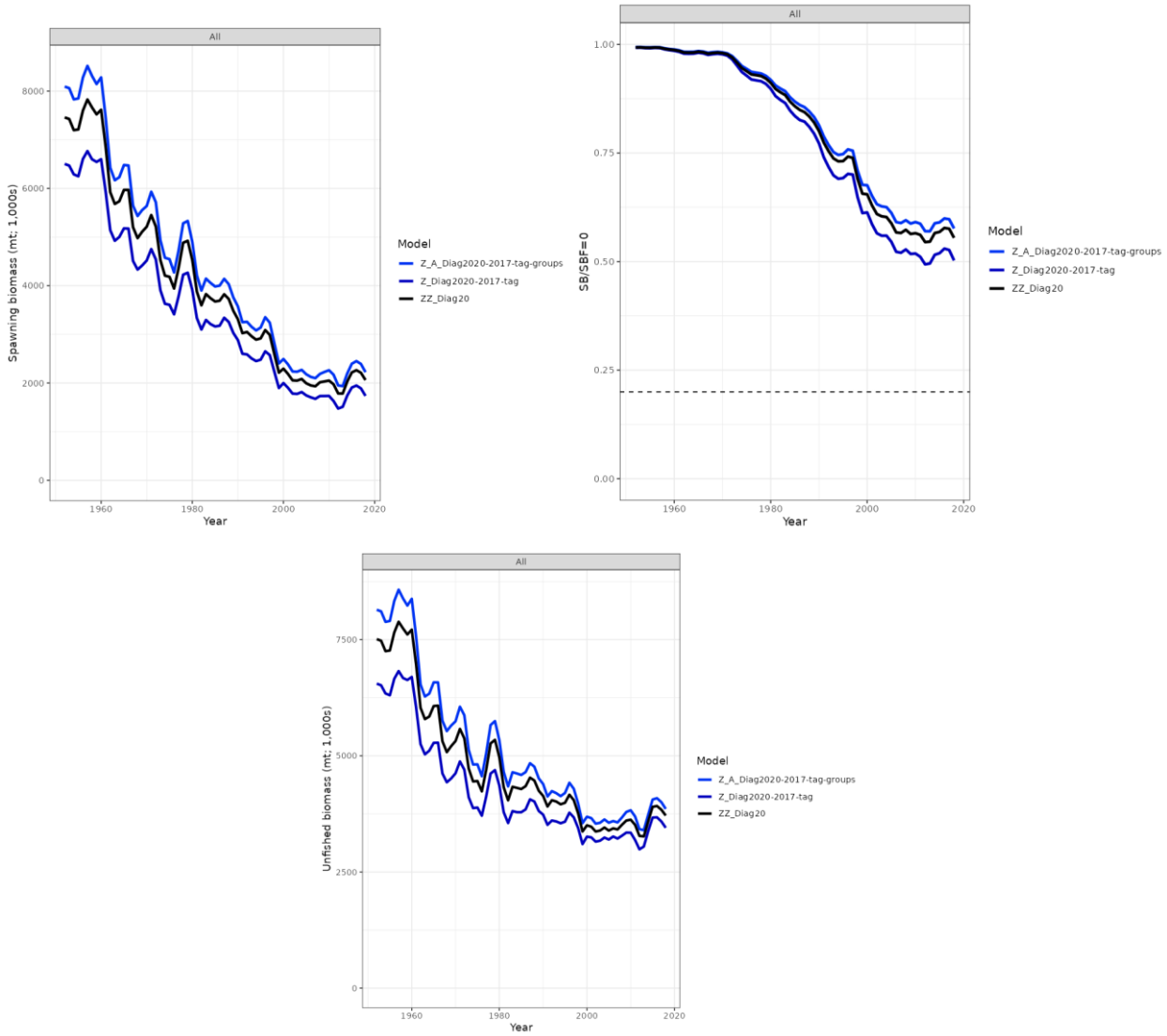


Figure 1(e). The change from the 2017 to the 2020 tagging data set led to more optimistic results. While several changes, including the addition of new data, were made to the tag-recapture data, it was found that using the 2017 tagging data in the 2020 diagnostic model led to a more pessimistic depletion outcome, and using the 2020 tagging data, but restricting it to the tag release groups available for the 2017 assessment produced a more optimistic depletion outcome

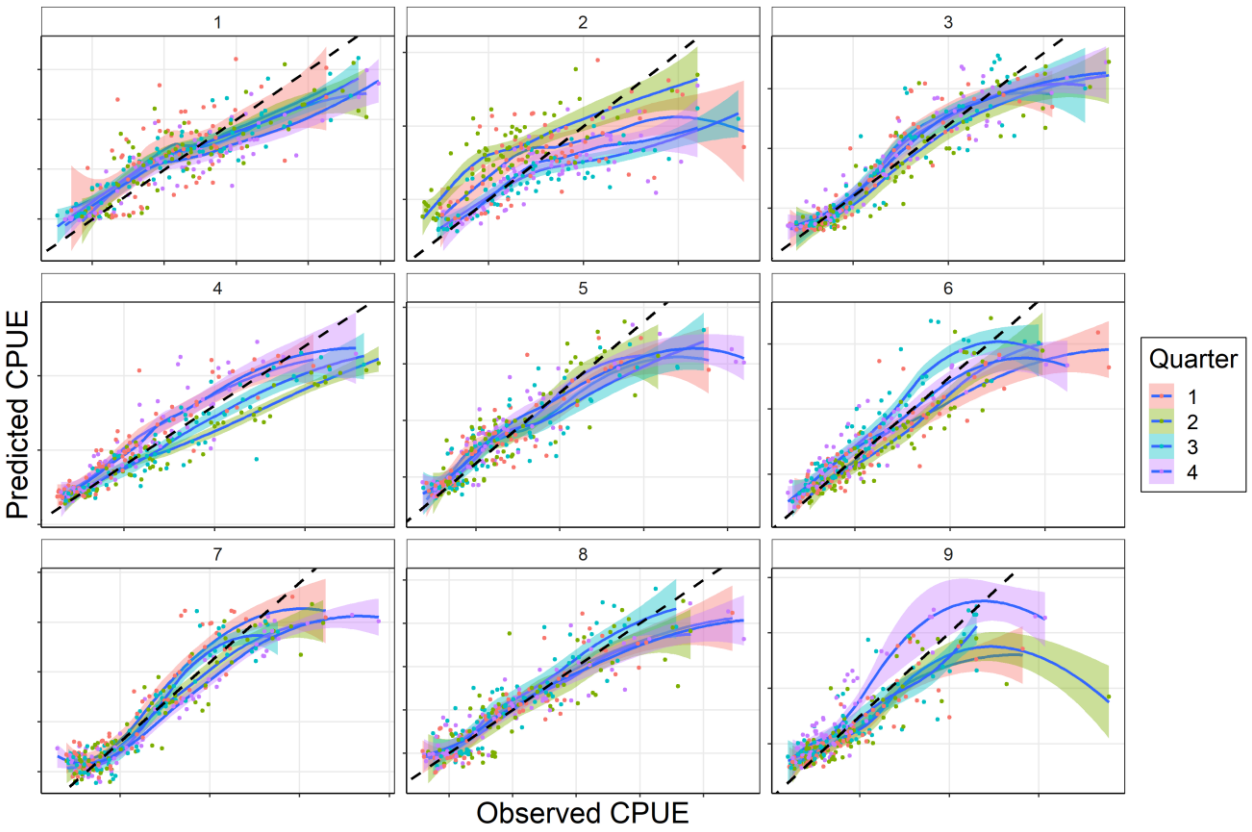


Figure 2. Observed vs model-predicted CPUE by region from the 2020 diagnostic model. with loess smoothers for quarter.

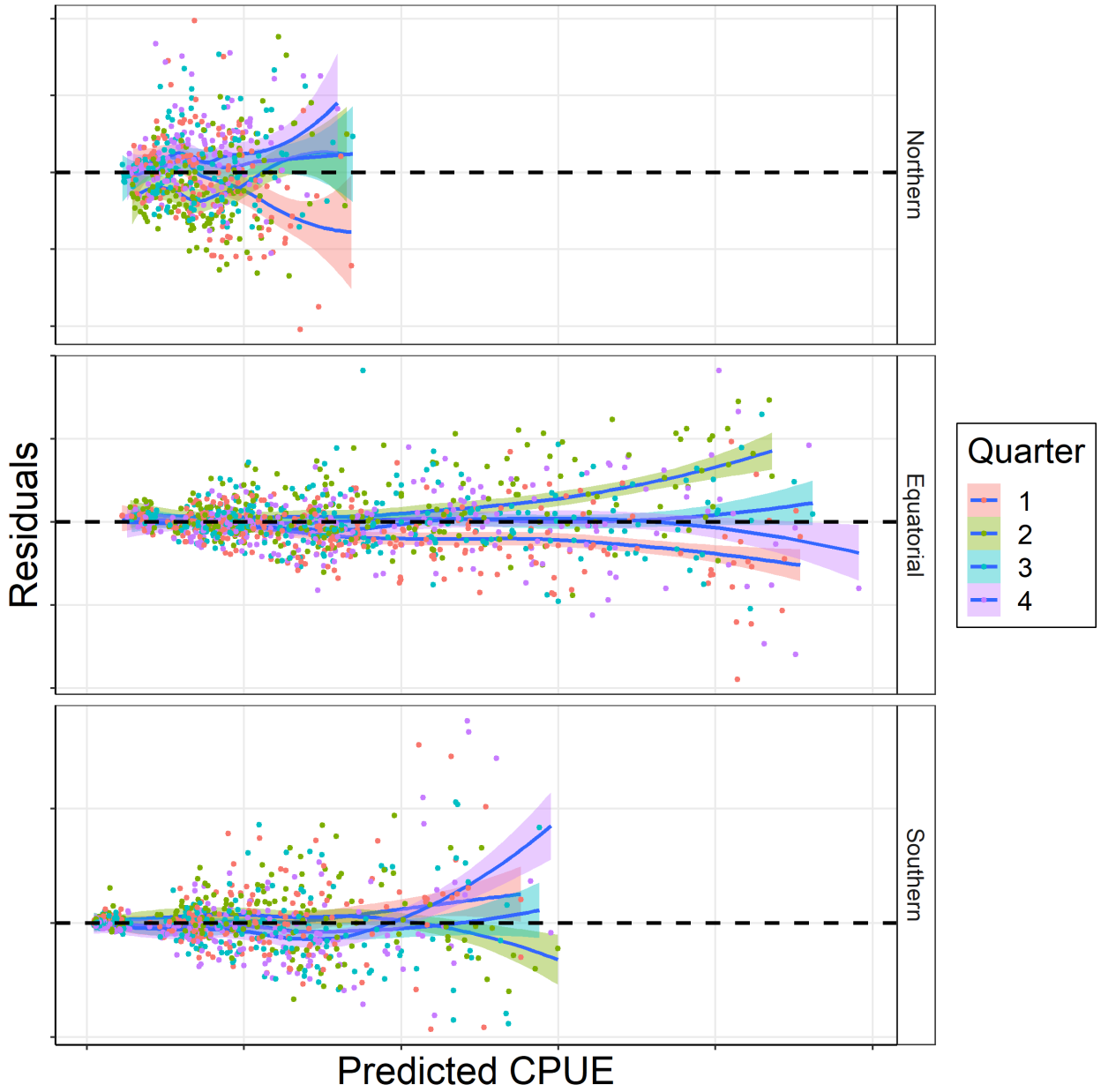
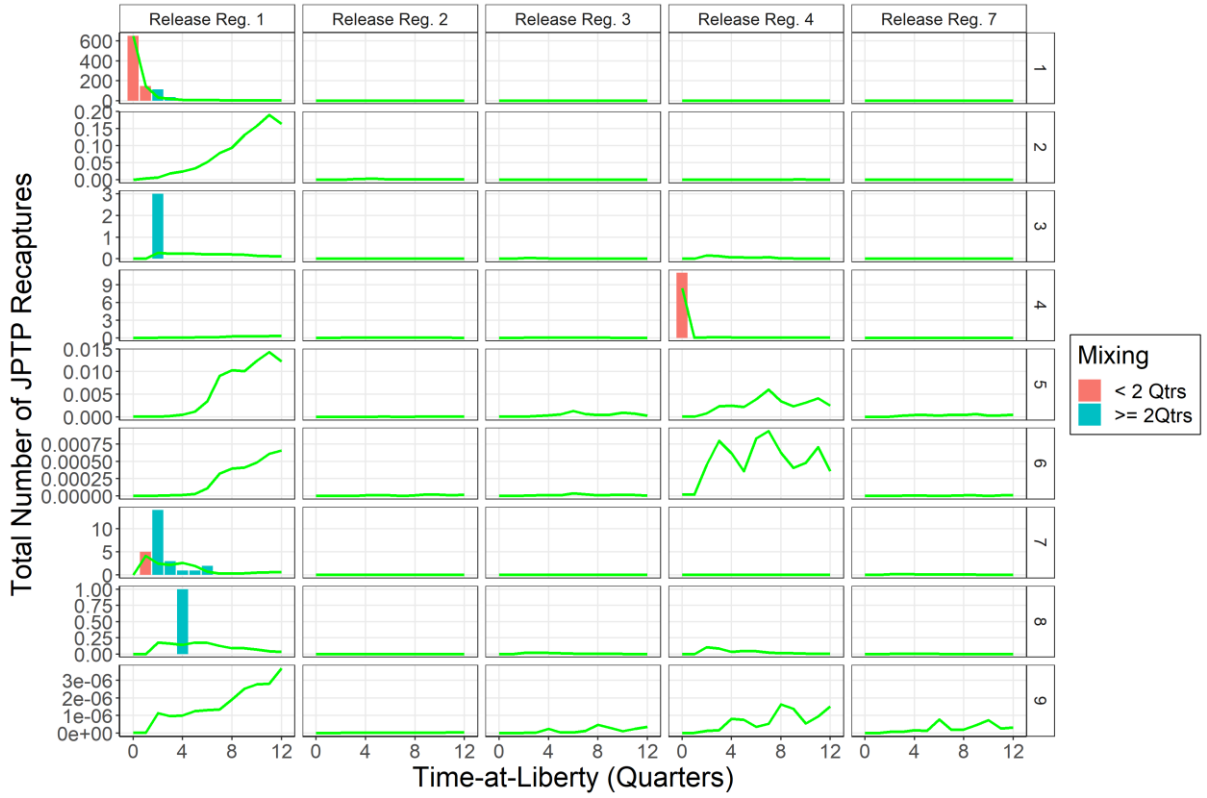
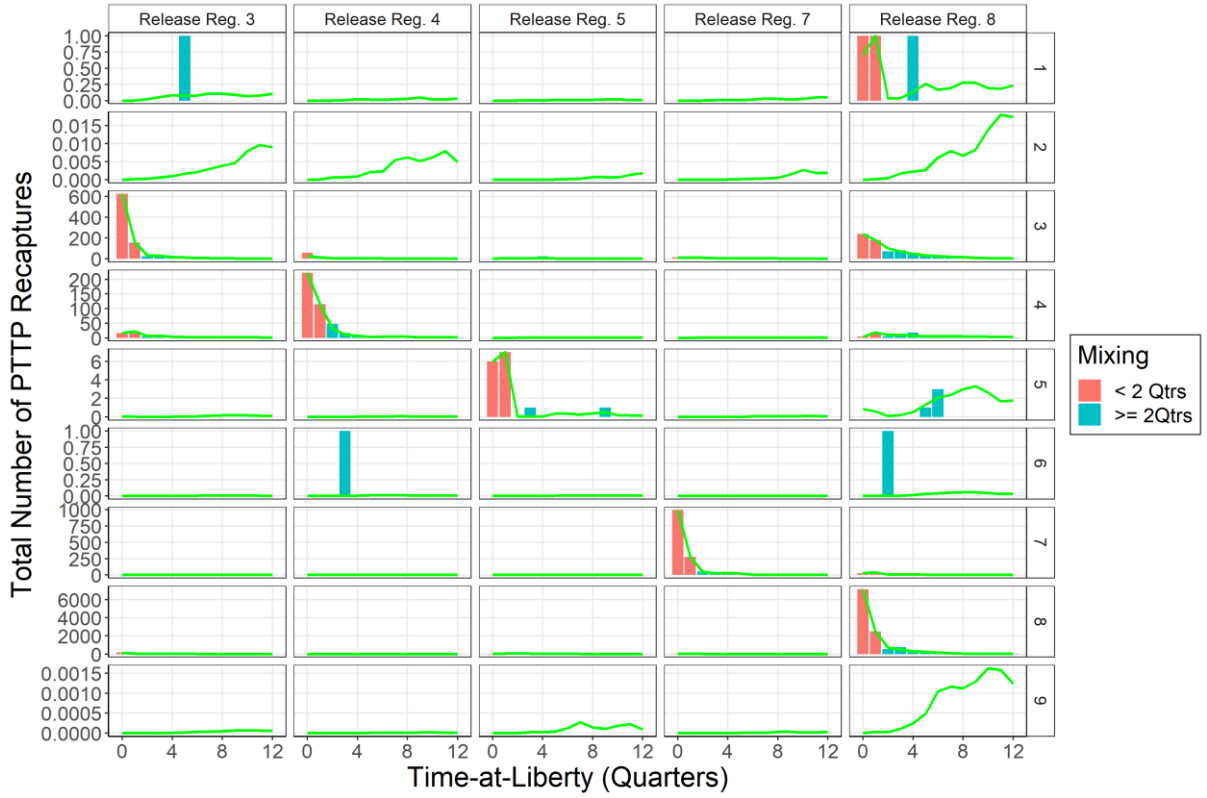


Figure 3. Observed vs model-predicted CPUE by three broad spatial areas from the 2020 diagnostic model. with loess smoothers for quarter.


Release Reg. 1
Release Reg. 2
Release Reg. 3
Release Reg. 4
Release Reg. 7
Total Number of JPTP Recaptures
Time-at-Liberty (Quarters)
Mixing
< 2 Qtrs
>= 2Qtrs


Release Reg. 3
Release Reg. 4
Release Reg. 5
Release Reg. 7
Release Reg. 8
Total Number of PTTP Recaptures
Time-at-Liberty (Quarters)
Mixing
< 2 Qtrs
>= 2Qtrs

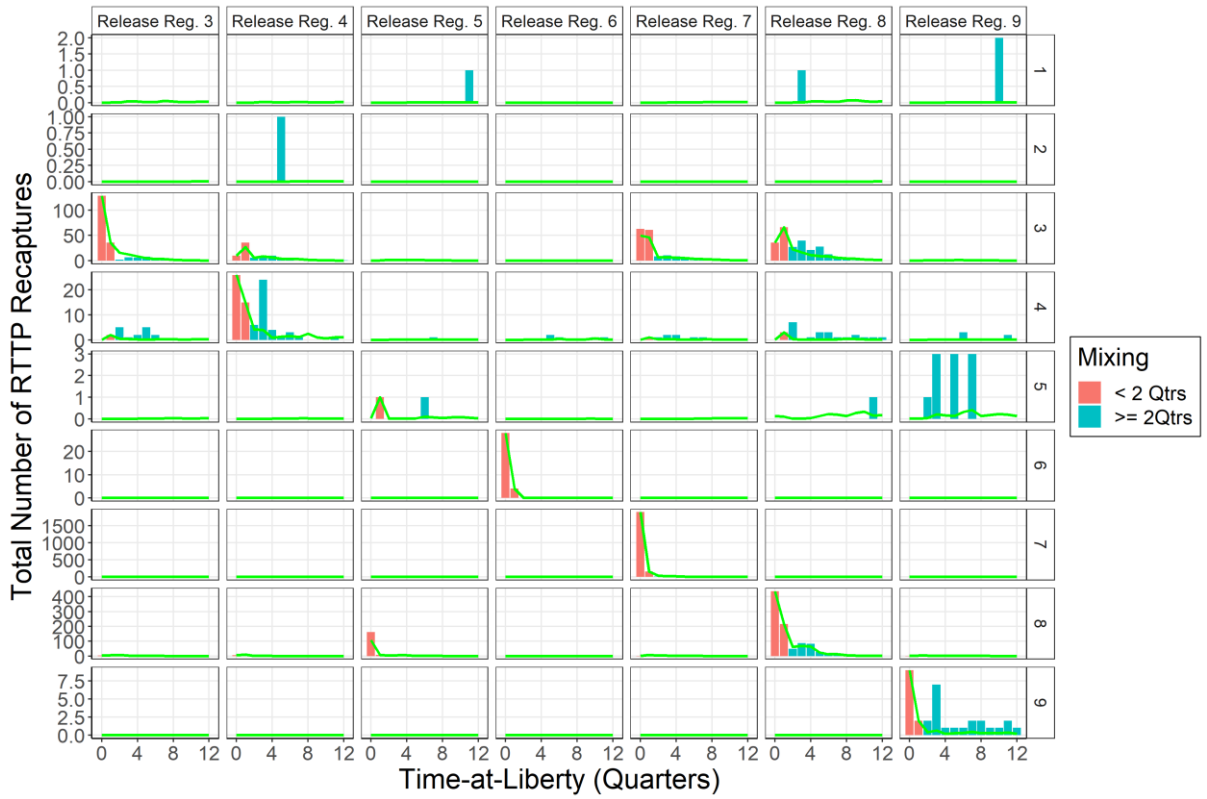


Figure 4. Example of a tagging diagnostic plot, with rows corresponding to recapture region

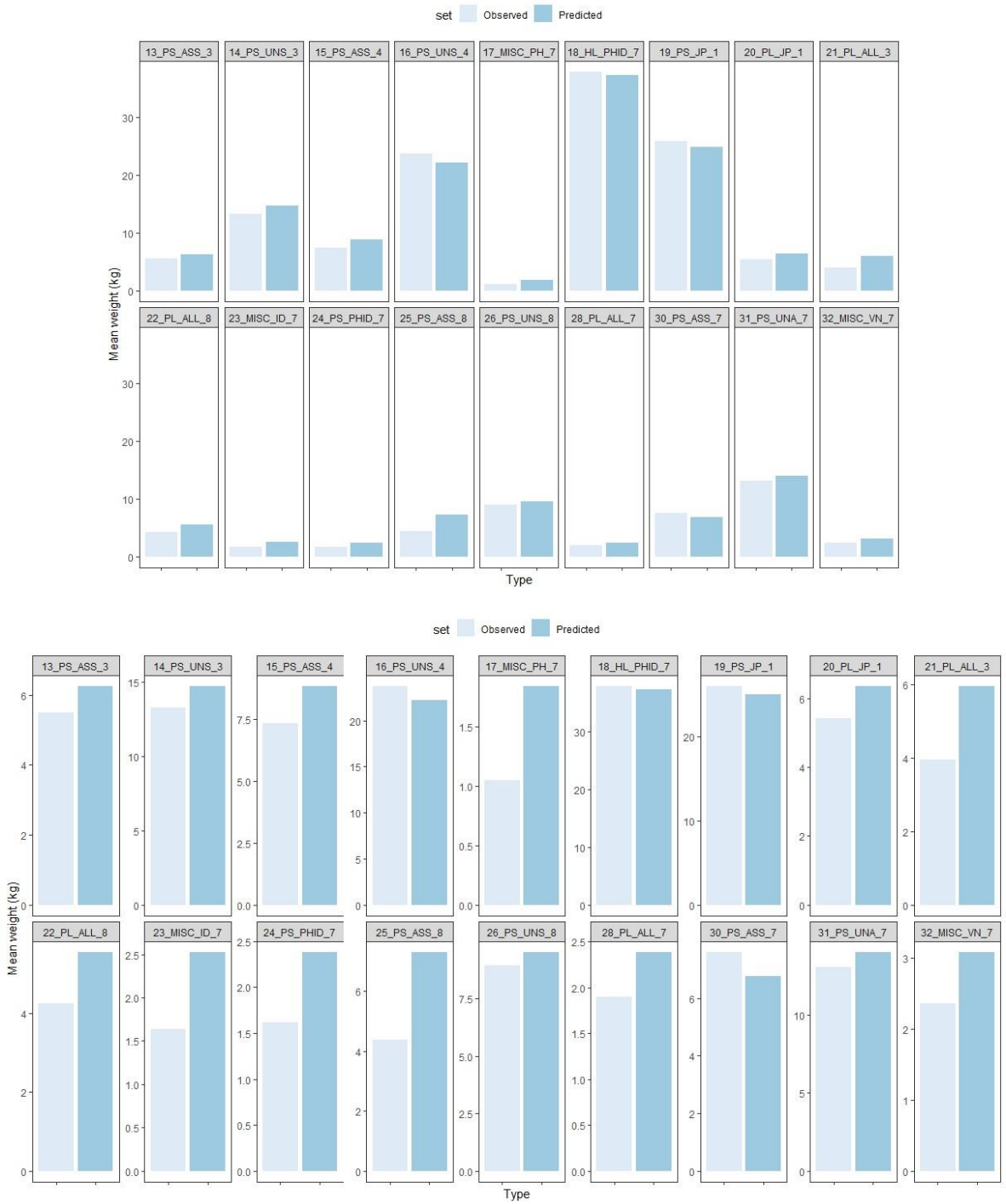


Figure 5. Observed and model-predicted mean weights for the fishery fleets where size compositions were measured by length in the 2020 diagnostic model. Top has constant Y axis scale.

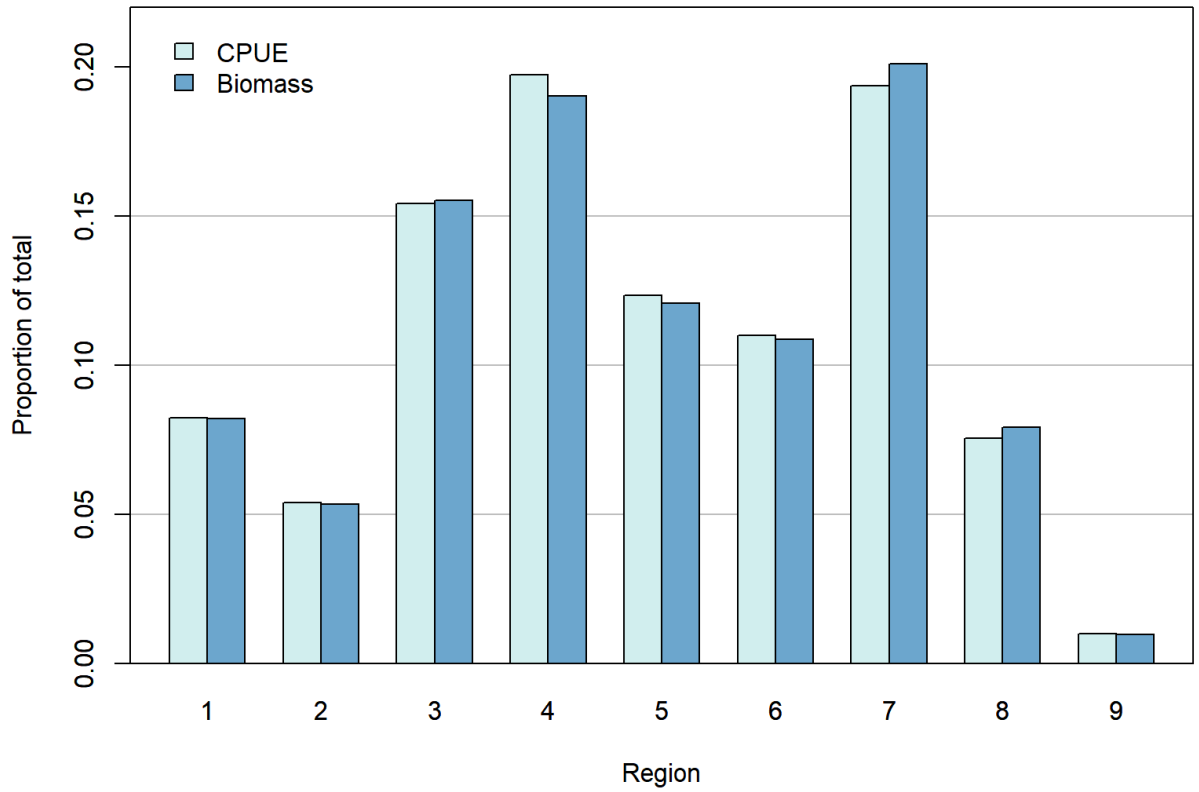


Figure 6. Average spawning potential vs average CPUE for the 2020 diagnostic model.

Appendix A: Panel Biographies

André E. Punt is a Professor in (and past Director of) the School of Aquatic and Fishery Sciences at the University of Washington. He received his BSc, MSc and PhD in Applied Mathematics at the University of Cape Town. Before joining the University of Washington, Dr. Punt was a Principal Research Scientist with the CSIRO Division of Marine and Atmospheric Research. His research interests include the development and application of fisheries stock assessment techniques, bioeconomic modelling, and the evaluation of the performance of stock assessment methods and harvest control rules using the Management Strategy Evaluation approach. He has published over 400 papers in the peer-reviewed literature, along with over 400 technical reports. Dr. Punt is currently a member of the Scientific and Statistical Committee of the Pacific Fishery Management Council, the Crab PLAN Team of the North Pacific Fishery Management Council, the Scientific Committee of the International Whaling Commission, and the IUCN Red List Standard and Petitions Committee. He is chair of the FIMS council and the Editor-in-Chief for the journal *Fisheries Research* and Associate Editor of the journals, *Population Ecology*, *Fishery Bulletin* and the *African Journal of Marine Science*.

Mark Maunder is the Head of the Stock Assessment Program at the Inter-American Tropical Tuna Commission. He received his B.Sc (Zoology and Computer Science), M.Sc (Zoology) at the University of Auckland and Ph.D. (Fisheries) at the University of Washington. Before joining the IATTC, Dr Maunder was a Quantitative Fisheries Scientist at the New Zealand Fishing Industry Board. His research interests include development of statistical methodology for fisheries stock assessment, protected species, and ecological modeling. He has coauthored over 100 papers in the peer- reviewed literature, along with many technical reports. Dr Maunder was co-founder and past president of the AD Model Builder Foundation, was a member of the Partnership for Mid-Atlantic Fisheries Science (PMAFS) Science Advisory Committee, and is Council Member of the Fisheries Integrated Modeling System (FIMS). Mark and his colleagues have been involved in extensive research into the development and application of fisheries stock assessment models. He was an early pioneer and advocate of the integrated assessment approach to stock assessment. His Phd dissertation involved integrating tagging data into stock assessment models. He was also the lead programmer of the general stock assessment model Coleraine that was an early ADMB based general model, and extensively used the integrated approach in a Bayesian framework, and codeveloped ASCALA that was used for assessing tunas in the EPO. He has also applied integrated analysis to protected species. In 2012, Mark co-founded the Center for the Advancement of Population Assessment Methodology (CAPAM; http://capamresearch.org/ [capamresearch.org]). The main activities of CAPAM revolve around the workshop series and associated special issues in the journal *Fisheries Research.* Mark has co-organized all the CAPAM workshops and chaired most of them. He has also been a guest editor for all the special issues is an Editorial Board Member with Fisheries Research. CAPAM has built an excellent reputation over the time it has been in existence, which has been recognized through being awarded the American Fisheries Society's (AFS) William E. Ricker Resource Conservation Award in 2018 and the American Institute of Fishery Research Biologists' (AIFRB) Outstanding Group Achievement Award in 2017.

James Ianelli is an affiliate professor at the University of Washington and a senior scientist with the NOAA's Resource Ecology and Fisheries Management Division of the Alaska Fisheries Science Center. Jim began his fisheries career as a tagging technician for the SPC tuna tagging program in 1979 and operated throughout the Pacific. He then worked for the IATTC in the mid-1980s on tuna tagging programs off of Mexico and in fishing villages in Japan. He became lab director of IATTC's Achotines facility before returning to work at SPC for the tuna program from 1985-1989. He then returned to the US and earned his PhD in Fisheries Science at the University of Washington in 1993. Since then, he continues to produce annual stock assessments for a number of important groundfish species in the North Pacific. His research interests include developing statistical approaches for ecosystem/fisheries conservation management. In addition to chairing the Scientific Committee of the South Pacific Regional Fisheries Management Organization and the North Pacific Fishery Management Council's groundfish Plan Team, he serves on the Advisory Panel for the Commission for the Conservation of Southern Bluefin Tuna.

Appendix B: Terms of Reference for an Independent Peer Review of the 2020 WCPO Yellowfin Tuna Assessment

Introduction

The 2020 yellowfin tuna (YFT) assessment (Vincent et al. 2020) in the WCPO (Western and Central Pacific Ocean) conducted by SPC using the MULTIFAN-CL assessment software was accepted by SC16 as the 'best available science' to inform managers of stock status. However, SPC noted that areas of uncertainty in the assessment required follow up investigation and expert advice, and that the assessment outcomes might provide an overly optimistic perception of stock status and the impact of fishing. SC16 recommended that follow-up work, including an independent peer review, was important to improve confidence in future YFT assessments for the WCPO. Given the similarities in model structure and data inputs, the follow-up work and peer review of the YFT assessment would also be relevant to the BET assessment (Ducharme-Barth et al. 2020).

This paper outlines a TOR for the peer review of the YFT assessment to be considered by SC17, which will guide the external review panel in their work. See Appendix 1 for the relevant extract relating to the SC16 recommendation for this peer review and suggested timelines.

This TOR provides the objectives and scope for the peer review. The process for running peer reviews of WCPFC stock assessments is outlined in the WCPFCs guidelines from SC12: Process for the Independent Review of stock assessments (Attachment K).

Background

The 2020 YFT assessment, beyond the addition of three years of tagging, catch, effort and size composition data, involved some notable changes from the previous assessments, namely:

- The implementation of the index fishery approach that used the geospatial (VAST) approach for CPUE standardisation
- Changes to how size composition data were prepared/reweighted
- Changes to the tagging data treatment
- Incorporation of new growth data from otoliths

Of these changes the tag mixing period, new growth estimation, selectivity assumptions, and data weighting appeared to have notable influence on the estimation of the key management quantities. The stock assessment indicated a more optimistic level of biomass and depletion than the previous assessments. A key concern, however, was that there was conflict among data sources in this assessment and depending on the amount of weight placed of different data sources, estimates of key management quantities could be quite different. The model structure may have also been overly complex given the available data and biological information. Further considerations post-assessment identified a number of areas related to input data, model structure and estimation approaches where follow-up investigations and advice were warranted. These considerations form the basis for the scope of this review.

Objectives

1. Undertake, in consultation with the stock assessment team (SPC), following the guidelines described in Process for the Independent Review of stock assessments (Attachment K), a peer review of the 2020 YFT stock assessment in the Western and Central Pacific Ocean (WCPO).
2. Based on the review work provide recommendations for improving the assessment, including data inputs, modelling approaches and treatment of uncertainty.
3. In conjunction with the SPC assessment scientists, identify improvement options that are feasible for application to the 2023 YFT assessment.

Scope

The key areas for consideration by the peer review panel based on the recommendations of the stock assessment report and follow-up considerations of the assessment team are listed below:

1. Model inputs, commenting on the adequacy and appropriateness of data sources and data inputs to the stock assessment, with particular attention to:
    a. **Growth:** review the approach to estimation of growth parameters and consider the implications of potential regional variations in growth.
    b. Tagging data: review the approach used to treat tagging data as model inputs, and how the tagging data are used within the modelling.
    c. Size composition: review the approach for pre-treatment of size composition data (i.e., re-weighting) and how size composition is weighted for the likelihood function.
    d. Natural mortality: review the approach used to determine M-at-age and implications of alternative M assumptions.
    e. **Data inputs:** identify and provide recommendations on the key areas for improvement in data collection (both fishery data and biological information).
2. Model configuration, assumptions and settings, with particular attention to:
    a. **Model complexity**: review the appropriateness of the model complexity, including spatial and fishery structure, in relation to data inputs and other available information.
    b. **Selectivity**: review selectivity assumptions and settings.
    c. **Uncertainty:** review the approach used to represent uncertainty in model-derived management quantities, considering structural, model and input data uncertainty.
3. Model diagnostics, with particular attention to:
    a. Review the suitability of the diagnostics used and reported for the assessment.
    b. Consider the diagnostics provided for the 2020 YFT assessment and provide guidance on follow-up work where the diagnostics suggest issues, i.e., data conflicts.
4. Recent MULTIFAN-CL model developments, with particular attention to:
    a. new MULTIFAN-CL features in relation to their application to the 2023 scheduled YFT assessment.
5. Future research areas, with the identification of priorities to improve future assessments.

While these key topics will be a focus of the peer review, other aspects of the assessment and data inputs may become focus areas as the review progresses.

Key activities and outputs from the peer review

Activity	Output	Timeframe
Review of the 2020 WCPO yellowfin stock assessment report	Summary paper of general comments and suggestions for any pre-workshop modelling or further information/data required by the review panel	To be provided by SPC to the panel by Jan 31, 2022.
Pre-workshop planning meeting. (Online)	Plan for the in-person workshop developed	At least 1 month prior to the September workshop.
In-person modelling workshop at SPC, Noumea	Completion of 5 day + travel in-person modelling workshop in Noumea	Planned date for this workshop is from 5-12th September 2022.
Review outcomes of modelling workshop	Draft workshop report to SPC	With 2 weeks of the end of the in-person modelling workshop.
SPC review of draft report	Draft report with any additional responses of SPC	The panel report with SPC comments is expected by mid-November 2022 and would align with any SC special session if this is requested by SC18.
Final report	Deliver report to WCPFC for posting	Final report be delivered to the WCPFC in February 2023 ahead of the SPC 2023 pre-assessment workshop in March/April. Final report to be discussed at the pre-assessment workshop to inform 2023 assessment.

The panel

The peer review panel was selected based on a CCM voting process co-ordinated by the secretariat, and is:
Dr André Punt – University of Washington
Dr Jim Ianelli – NOAA
Dr Mark Maunder – Inter-American Tropical Tuna Commission

Logistics and COVID implications

The expectation following SC16 was that the review would commence at the start of 2022 with the review reporting to SC18 (August 2022) and informing development of the 2023 YFT assessment. The SPCassessment team, including people involved in the previous peer review of the BET assessment, and expressions of some CCMs at SC17, indicate a strong preference for an in-person workshop in Noumea to be part of the review process. The peer review is unlikely to be successful without the free discussion and adaptability of an in-person modelling workshop. The issue of time differences also makes working online in a flexible, interactive and adaptive way very difficult. The uncertainty of the COVID-19 situation and travel options means that timing of the Noumea workshop will need to be flexible. There is also the issue of the requirement for a quarantine period (currently 7 days for arrivals in New Caledonia if vaccinated). This is not ideal given the busy schedules of the review panel and the SPC assessment staff. It now appears that the workshop will not be feasible before SC18 and will need to be scheduled later in 2022, SPC suggested two options for the SC17 to consider:

1. The review report be presented to the SPC Pre-Assessment Workshop in 2023 to provide the opportunity for CCMs to comment and discuss recommendations and approaches to consider for the 2023 YFT (and BET) assessment. In that case the review report would be formally submitted to SC19 as a supporting document for the 2023 assessment.
2. Submit the review report some time after SC18 for intersessional consideration, either through an 'Online Discussion Forum' or an online meeting, or potentially both. Submit the revised report, with responses to comments, to the Secretariat for posting, and then present an overview of the review findings and recommendations at the SPC Pre-Assessment Workshop in 2023.

Note: SC17 (summary report paras 242-245) did not provide any specific recommendations on the process for delivery of the peer review outcomes but some CCMs noted strong preference that the review include an in-person workshop between SPC and the peer review panel, and others expressed support for Option 2 including to have a 1-2 day special SC session before work on the 2023 stock assessments commences.

While the in-person workshop is a key part of the review process, it is now also planned to have approximately 3 monthly meetings between the peer review panel and SPC staff to discuss and set modelling tasks and review results of previous tasks. In this way the review work can progress in an iterative fashion and not be totally dependent on the work shop.

References

Ducharme-Barth, N., Vincent, M., Hampton, J., Hamer, P., Williams P., and Pilling, G. (2020). Stock assessment of bigeye tuna in the western and central Pacific Ocean.WCPFC-SC16-2020/SA-WP- 03(REV3)

Vincent, M., Ducharme-Barth, N., Hamer, P., Hampton, J., Williams P., and Pilling, G. (2020). Stock assessment of yellowfin tuna in the western and central Pacific Ocean. WCPFC-SC16-2020/SA-WP-04-Rev2

Appendix 1

Relevant extract from SC16 "Outcomes Document https://www.wcpfc.int/node/47653

3.6.2 Peer Review Recommendations

70. SC16 supports an external expert peer review of the yellowfin stock assessment. This would also allow several components of the bigeye tuna assessment to be reviewed given the similar data input structure. This review would examine a number of issues such as model complexity, weighting of data sources, spatial approaches and the extreme sensitivity to assumptions on growth amongst a range of other issues.

71. SC16 provides the following provisional time-line for an external expert peer review.

    i) Year 1 would be set aside to allow the SSP to conduct an initial range of testing and analysis internally focussed on YFT and report these findings to SC17. SC17 to finalize ToRs for the external expert review.

    ii) Year 2 would be set aside for the SSP to conduct further testing and analysis internally focussed on BET and YFT, following SC17 input, and for the external expert review (commencing at the start of 2022) with the review reporting to SC18.

    iii) Year 3 would provide updated YFT and BET stock assessments which respond to the review. The two assessments would be reported to SC19.

72. In accordance with this, SC16 identified the external review as a project in the budget (provisionally estimated at $USD 50,000) but with no funding commitment until 2022 and 2023.

73. SC16 also tasked the SSP with preparing a draft terms of reference for the external expert review for the consideration of SC17 which would be informed by their analyses during 2021. The draft terms of reference would give consideration to including the bigeye stock assessment in the external review process.

74. Further, SC16 noted that peer review experts of the required calibre may not be easy to secure, thus efforts should be made during late 2020/early 2021 to have them express interest

Appendix C: Issues identified for the review meeting

A. Comparing 2017 and 2020 assessment

1. The 2020 assessment estimates a more optimistic stock status than the 2017 assessment. Is this because of a larger numerator (SB), smaller denominator (SBF=0, dynamic B0), or both?
2. The 2020 assessment estimates a more optimistic stock status than the 2017 assessment. Is this because of changes in recruitment, M, maturity, body weights, F, or dynamic B0 between the two assessments?
3. Of all the changes between the 2017 and 2020 assessments, which ones had the greatest effect on the estimated stock status?

B. Selectivity

1. Which fisheries, if any, should be grouped in terms of selectivity?
2. Should one longline fishery within each region have a non-decreasing selectivity?
3. Should selectivity be modelled using cubic splines or parametric curves?
4. Selectivity can only be modelled as age-based in MFCL, is this problematic for the YFT assessment?
5. Are the fits to the purse seine length compositions adequate, or can selectivity be improved for the purse seine?

C. Growth

1. Should the growth curve be estimated internally in the assessment model or externally?
2. Should lengths from tagging data be included when fitting an external growth model?
3. Should the growth follow a parametric von Bertalanffy curve, a nonparametric curve, or in between (e.g., first 8 ages nonparametric)?
4. Should other growth curves be considered as alternatives to von Bertalanffy?
5. Should the external von Bertalanffy growth curve model be fitted to otoliths only or tagging data plus otoliths? Results from both analyses are available, presented in 2020.

D. Maximum age

1. What is an appropriate maximum age in the YFT assessment?

E. Tags

1. Should the tag mixing period be 2 quarters or tag release group specific (as done in the SKJ 2022 assessment)?
2. Is the tagging data informative about migrations, mortality, and/or stock size?
3. Could the tag-related plots and tables in the assessment report be improved?

F. Regions

1. Is it worth considering a simpler regional structure, e.g., a 4-region model which might capture the main dynamics of the YFT fishery and have better statistical properties of estimability?
2. What should be the basis of regional boundaries? Aspects to consider, including biology, fishery, management, model parsimony, model convergence, estimability, ease of diagnostics, ease of interpretation?
3. Do the data indicate a change in the distribution and/or movement that could be due to climate change, as has been the case with some other fish stocks in the region, and is there something that could be improved in the assessment to handle such changes?

G. New MFCL features

1. Should fishing mortality be estimated using a catch-errors or catch-conditioning approach?
2. Should recruitment be estimated using an orthogonal-polynomial approach?
3. In addition to the Dirichlet-multinomial approach to weight length compositions, should the uncertainty grid include arbitrary sample size scalars?

H. Natural mortality

1. Review the approach used to determine M at age and implications of alternative M assumptions. Could tagging data be informative for estimating M?
2. What shape would be appropriate when estimating M at age from life history parameters?
3. What effect did the change in M have on depletion?

I. Data collection

1. Are there gaps in the data collection that could be improved in the sampling programme?
2. What fisheries and regions should be sampled for future close-kin mark-recapture data collection?
3. What sex-specific data do we have and how are/could they be used?

J. CPUE

1. Should effort creep be included in the CPUE data and what is the best way to do that?
2. Can the VAST analysis from 2020 of yellowfin longline CPUE data be improved for the next assessment?
3. Should multiple CPUE indices be considered?
4. Should additional process error be added to the CPUE index, e.g., using a loess smoother and/or estimating an additional standard error?
5. Should CPUE catchability ungrouped or grouped between regions, enabling regional scaling?

K. Model complexity

1. Is the level of model complexity appropriate, including spatial and fishery structure, in relation to data inputs and other available information.
2. Are the patterns of high recruitment in temperate regions and zero recruitment in region 8 (PNG and Solomons) in agreement with the available data, or can the model be simplified to reduce the possibility of model balancing unrelated to data?
3. Would a sex-specific model be likely to be helpful for providing management advice?
4. Can the review panel provide technical recommendations to consider if the yellowfin stock was analyzed using an alternative model framework, such as Stock Synthesis?

L. Uncertainty

1. Is there a better way to represent uncertainty about management quantities, combining structural and estimation uncertainty?
2. What model runs should be included in the structural uncertainty grid, as opposed to one-off sensitivities, and how should they be weighted?
3. If the estimation uncertainty about depletion is very small, evaluated using the delta method or likelihood profile, what other approaches could be used to evaluate the estimation uncertainty?

M. Diagnostics

1. Which standard stock assessment plots would be useful but are not provided in the assessment report?
2. Which diagnostics would be useful but are not provided in the assessment report?
3. How should model convergence be addressed in the assessment report, criteria such as jittering of initial parameter values, parameters on bounds, final gradients, positive definite Hessian, parameter correlations, etc.
4. Is there consistency between input and output variances for the CPUE, length compositions, and tag recaptures?
5. Diagnose possible problems fitting the data in Fishery 18, Indonesian handline in Region 7. Is the observed catch in tonnes higher than the estimated exploitable biomass, and is the fishing mortality hitting a parameter bound at 1.3?

N. Summary plots and tables

1. Data: Total length comps over time (bubble plot or 3d histograms)
2. Data: Median length over time, with confidence limits and possibly overlaid with (Results) a line showing median length in the model population?
3. Results: Population numbers at age as a table and/or bubble plot?
4. Results: Also plot CPUE by year instead of quarters?

Appendix D: Background documents considered by the Panel

- Ducharme-Barth, N. and M. Vincent. Analysis of Pacific-wide operational longline dataset for bigeye and yellowfin tuna catch-per-unit-effort (CPUE). WCPFC-SC16-2020/SA-IP-07
- Hoyle, S., Nicol, S. and D. Itano. 2009. Revised biological parameter estimates for application in yellowfin stock assessment. WCPFC-SC5-2009/BI-WP-3-rev 2.
- McKechnie, S., Harley, S., Davies, N., Rice, J., Hampton, J. and A. Berger. 2014. Basis regional structures used in the 2014 tropical tuna assessments, including regional weights. WCPFC-SC10-2014/SA-IP-02.
- Scutt Phillips, J., Lehodey, J., Hampton, J., Senina, I., and S. Nicol. 2022. Quantifying rates of mixing in tagged, WCPO skipjack tuna. Technical Report WCPFC-SC18-2022/SA-WP-04.
- Tremblay-Boyer, L., McKechnie, S., Pilling, G. and J. Hampton. 2017. Stock assessment of yellowfin tuna in the western and central Pacific Ocean. WCPFC-SC13-2017/SA-WP-06 (Rev 1).
- Vincent, M., Ducharme-Barth, N., Hamer, P., Hampton, J., Williams, P. and G. Pilling. 2020a. Stock assessment of yellowfn tuna in the western and central Pacific Ocean. WCPFC-SC16-2020/SA-WP-04 (Rev3).
- Vincent, M., Ducharme-Barth, N. and P. Hamer. 2020b. Background analyses for the 2020 stock assessments of bigeye and yellowfin tuna. WCPFC-SC16-2020/SA-IP-06.

Appendix E: Participants during the review

Review Team
Jim Ianelli: NOAA, NMFS
Mark Maunder: IATTC
André Punt: University of Washington

SPC
John Hampton
Nick Davies
Arni Magnusson
Jemery Day
Paul Hamer
Claudio Castillo Jordan
Thom Teears
Nan Yao
Sam McKechnie
Simon Nicol

Appendix F: Requests to the Analysts

#	Request	Rationale	Response
A	Provide assessment model outcomes in the form of time-trajectories of spawning potential (SSB), unfished SSB (SSB_F=0) and the ratio of spawning potential to unfished spawning potential.	Spawning potential relative to unfished SSB is a key model output but the time-series of spawning potential and SSB_F=0 differ in unexpected ways and the time-series of SSB_F=0 often differs more among past assessments than SSB. The reasons for this are explored in several of the requests.	This was completed and is now a standard output in the ShinyApp.
B	Create a table of coefficients of variation for the CPUE series.	The Panel wished to better understand how the CPUE indices are weighted in the model fitting process. It should be noted that understanding the variation of CVs among regions is important because the catchability and selectivity are assumed to be shared.	A plot of CVs over time was provided, which showed higher CVs during the earlier years and particularly for region 9. The CVs for region 3 were the lowest (~0.1) so the model places greatest weight on fitting the CPUE index for this region.
C	Document the equations used when fitting the CPUE data.	CPUE is a key data source in the assessment and the Panel wished to better understand the various factors involved in how these data are used in the assessment.	This was covered in Dr. Nick Davies' presentation during discussion of the new CPUE likelihood in MULTIFAN-CL.
D	Plot (for the 2020 diagnostic model) the observed versus model-predicted CPUE values. Color each season differently.	The Panel wished to better understand whether the model fits the data adequately, and whether a quarterly catchability coefficient should be applied or if quarterly movement is adequate.	Figures 2 and 3 plot the model-predicted CPUE values versus the residuals, with loess smoothers for quarter. There are some patterns (e.g., regions where fits are poorer). This could be due to data conflicts and should be accounted for the future model development.
E	Document how grids with no observations are treated in the VAST, document how each polygon was weighted when computing the CPUE indices, and how the integration over grids was conducted, and explain how spatial correlation might impact the results.	The Panel was concerned that grids with no data (e.g., in the 1950s) might be ignored and grids equally weighted when computing the CPUE indices, which would lead to bias in the resulting CPUE indices. It was also concerned that densities for high density areas could be extrapolated to grids with little data.	The Panel was provided with information on how the grids with few or no data were dealt with. The indices were computed by summing the product of densities and area by 1^0x1^0 cell, where the area was set to actual area of the cell. The 1^0x1^0 cell density is down-projected from the estimated density at each spatial knot and time step (e.g., for any given time-step all 1x1 spatial cells associated with the same knot will have the same predicted density).
F	Provide plots of the SSB and recruitment separately for the north, equatorial and south areas for the 2020 diagnostic model	The Panel wished to better understand the trends in relative abundance.	Figure F.1 shows that the depletion differs spatially and by 2020, with about half of the biomass in the north and south areas.
G	Document the basis for the regions and the rationale for the selected boundaries.	The Panel understood that there was a desire to maintain consistency between the regions for the bigeye and yellowfin assessments, but also wished to confirm that the selected regions can be justified based on data analyses.	McKechnie et al. (2014) was provided that outlines the basis for the regions and their borders. This is discussed further in section B.5.
H	The length-weight regression parameters were updated as part of the 2020 assessment. Was this to address spatial and temporal variation in these parameters?	Bias in the estimates of these parameters could be very consequential for the results of the assessment, particularly since the model fits to weight frequency data.	The length-weight regressions were updating by adding data for smaller animals. Modifying the 2020 diagnostic model to use the old length-weight parameter values suggested little sensitivity to the values of these parameters (Figure 1).
I	Update the plot examining whether the growth curve is consistent with the length-frequency data (e.g., by adding the model-predicted length-frequencies, and highlighting cohorts).	One of the major reasons for changes to the results of the assessment from 2017 to 2020 is how growth is treated, but it was unclear whether the resulting growth curve fitted the data adequately. It would be expected that the model should track the mean length-at-age for strong cohorts. There was also concern that growth might differ among regions and this might be apparent in misfits to the composition data.	Plots were provided that showed the weight-at-age distributions for the cohorts that make up the observed distributions of weight. This illustrated that the von Bertalanffy model with offsets led to offsets that appear to mimic the lack of growth for 40-50cm yellowfin. However, the fits were generally quite poor, including for the model that fitted to the data without the conditional age-at-length data. The MULIFAN-CL viewer was also provided to the Panel for use in investigating the issue.
J	Update the specifications for the model runs to reflect all the changes made to the model.	The analysts noted that some of the putative changes reflected changes to aspects of the model in addition to the stated change, and some of the undocumented changes may be influential.	Appendix G lists the specifications for each of the model runs in the bridging analysis.

#	Request	Rationale	Response
K	Create a table of parameters for the 2020 diagnostic model, indicating which are estimated versus being pre-specified.	The diagnostic model has many parameters, but which are estimated and which are pre-specified cannot be easily discerned from the documents provided.	The table was produced. It should be included routinely in assessment reports to enable an evaluation of how the number of estimated parameters changes with changes to model configuration
L	Document how many tagged animals were dropped from the analysis when the definition of "2 quarters" was changed to 182 days.	This change appears to have a major impact on the results of the assessment, but it is unclear how many tags are involved.	This request was completed. The analysis, however, highlighted that much of the impact of the associated change between models SelUngroup and JPTP may have involved increasing the number of releases owing to the method used to deal with tagging-induced tagging mortality. This is explored further in request CC.
M	Correct the plot of where biomass by region comes from which regions	The plot in the assessment report indicated that much of the biomass in region 8 originated in region 8, but this cannot be correct given there is (almost) no recruitment to region 8.	This request could not be completed during the review meeting, but should be addressed for the future assessment.
N	Explore the effects of changing the plus-group age and the parameters of the length-weight regression separately.	These two effects were changed at the same time in the bridging analysis and the Panel was interested to understand the effects of each change separately.	This analysis was conducted for the 2020 diagnostic model and indicated neither increasing the number of age-classes nor changing the length-weight parameters had a major impact on the results of the assessment (Figure 1). This suggests that a third aspect of Age10LW_SelStep model, changing *M*-at-age led to the marked change between the JPTP and Age10LW_SelStep models. The change in *M*-at-age was attributed to the use of a different sex ratio-at-age.
O	Repeat the CondAge model omitting the conditional age-at-length data.	The CondAge model involved several changes in addition to adding the conditional age-at-length data and the Panel was interested to explore the effect of just adding the conditional age-at-length data.	This run involved dropping the conditional age-at-length data but still using the von Bertalanffy growth curve that does not have any offset parameters. It did not lead to plausible outcomes, likely due to changes to several of the other specifications.
P	Document what was actually done in the CondAge model when the catchability parameters were changed and maturity-at-age was treated differently.	The changes to the model were unexpected and the Panel wished to better understand the changes actually made.	The 2017 reproductive output-at-age vector was based on maturity-at-age, fecundity-at-age, proportion spawning-at-age and sex-ratio-at-age. It was noted that the sex-ratio-at-age vector was updated for the 2020 assessment. This led to a request to better understand the basis for the 2020 reproductive output-at-age vector (Request R).
Q	Document the basis for the 2017 reproductive output-at-age vector.	This vector implies few females older than age 28, but the 2020 assessment indicates that many of the animals of age 28 and older should be females and hence contribute to reproductive output.	Including the 2017 reproduction output-at-age vector into the 2020 diagnostic model led to a more optimistic result (Figure F.2), owing to the reduced impact of older fish on spawning potential. Thus, the change to reproductive output-at-age vector does not appear to be reason for the more optimistic results, but is an influential difference from the 2017 assessment.
R	Document the basis for the 2020 reproductive output-at-age vector.	Insufficient information was available to fully understand how the vector was constructed.	The method is fully documented in Hoyle et al. (2009). It seems the "Full" sex ratios-at-age (Hoyle et al., 2009) were used in the 2011, 2014 and 2017 assessments, but the sex ratio-at-length data were changed to being based on the observer records for the 2020 assessment.
S	Run the 2020 diagnostic model with the 2017 *M*-at-age vector. Show spawning potential time-trajectories using the 2020 and 2017 reproductive output-at-age vectors.	The Panel wished to see the effects of changing the *M*-at-age vector on its own.	Running the model with the 2017 *M*-at-age led to a minor increase to the estimates of recruitment, as expected (Figure 1). Spawning potential is lower when the 2017 *M*-at-age vector is used. Overall, had the 2017 *M*-at-age and reproductive output-at-age vectors been used in the 2020 assessment, the results would have been more optimistic.
T	Run the 2020 diagnostic model assuming that growth is governed by the Richards curve.	The Richards curve model in the uncertainty grid was based on fitting this curve outside of the assessment, but the way the length-at-age data were collected likely led to a biased estimate of length-at-age.	The model crashed. This is addressed further in Request MM.
U	Evaluate the fits to the data on length-at-age: (a) on the map, plot points colored differently if the residual is positive or negative, and (b) on the growth curve plot the points colored by region.	The Panel wished to understand the evidence that growth differs spatially.	The results (Figure F.3) are suggestive of spatial difference in growth (faster growth in the temperate areas), but the data set remains quite small. Evaluation of the modes in the length composition using the MULTIFAN-CL viewer corroborated these findings

#	Request	Rationale	Response
V	Review the properties of the "discarded" otoliths (were they older within length-classes)	Some otoliths were discarded as unreadable, but they should be random with respect to age within a length-class to avoid bias.	Figure F.4 shows the proportion of otoliths rejected due to an inability to agree an age. The proportion is not independent of age, which could lead to bias when a growth model is fitted to conditional age-at-length data. Future assessments should explore sensitivity to including all of the age data irrespective of whether the ages are agreed and including age-reading error.
W	Compute the conditional age-at-length residuals (Pearson residuals within each length-class)	The residual plot in the assessment report did not plot the residuals about the fit to the conditional age-at-length data correctly.	The conditional age-at-length residuals (Figure F.5) are suggestive that the age of large animals is over-estimated. This occurs even though the conditional age-at-length data are fully weighted compared to the other data sources.
X	Estimate of *M*-at-age in the 2020 assessment	The Panel could not discern the exact algorithm used to calculate *M*-at-age from the assessment report.	The average *M* dropped from 0.23 to 0.2 yr^{-1}, but the Panel remains unconvinced that *M* can be estimated within the assessment.
Y	Run the 2020 diagnostic model with the 2017 tagging data. If the difference in final depletion differs by more than 0.05 from the original version of the 2020 diagnostic model then examine the effects of each change to the tagging data	The Panel wished to understand which change to the tagging data led to the change in estimated depletion.	Using the 2017 tagging data led to more pessimistic results (by 5 percentage points), with the estimates of regional biomass and hence depletion changing differently. Using the 2020 tagging data the JPTP data set and the recaptures led to more pessimistic results (Figure F.6).
Z	Construct a diagnostic plot for the tagging data that plots (for each release group) the numbers observed and predicted to be recaptured by region, with the totals (over time) observed and predicted to have been recaptured included for each region.	The Panel wished to better understand how well the model replicates the tagging data, and whether the model can replicate the number of tags that were recaptured in regions other than those in which they were released.	The plot was provided (Figure 4) and should be refined further and included in future assessments,
AA	Run the 2020 diagnostic model fixing the reporting rates to those based on tag seeding when estimates based on tag seeding are available and estimate the remainder.	Several of the estimates of the tag reporting rate differ markedly from their prior means and several tag reporting rates are found on boundaries. This request would remove the first of these concerns.	As expected, forcing the tag reporting rates to match the means of their priors led to less optimistic results.
BB	Run the 2020 diagnostic model setting the reporting rate to zero, excluding the recaptures, and adjusting the releases for the number of excluded recaptures for those regions/fisheries for which no data on tag reporting rate is available from tag seeding experiment.	The Panel was interested to understand the effect on the model results of tags recaptures in regions/fisheries with few data, but recognized that this may lead to an unstable model.	This run could not be completed during the review meeting given its complexity.
CC	Run the 2020 diagnostic model using the 2017 tagging data, but apply a 3-quarter mixing period.	There is no clear basis for the 2-quarter mixing period and the Panel wished to understand the consequences of a greater extent of exclusion.	The results are more optimistic than the run with a 2-quarter mixing period and the 2017 tagging data confirming that the tagging data are in conflict with the other data included in the assessment.
DD	Document how the aggregate length- and weight-frequency plots were calculated, confirm that the model-predictions for each quarter were weighted by the sample size for that quarter.	The Panel wished to better understand how the plots were constructed, and that quarters with vastly different sample sizes were not equally weighted when computing the model predictions.	The observed and model-predicted length- and weight-frequency data are correctly weighted in these plots.
EE	Calculate the mean weight for each observed and predicted composition by fishery (from the length-frequency data) to assess if the right number were being removed given the model fits the catch weight well.	Some of the length-frequency fits are poor and may result in the "wrong" number of animals being removed from the population.	The plot was produced (Figure 5) and indicated that the model overestimates the mean weight fisheries with a mean catch weight of 15 or less kg and vice versa. It was noted that some of these are large fisheries and plots like Figure 5 should be routinely included in assessment reports.

#	Request	Rationale	Response
FF	Document how the spline selectivities were specified (e.g., how were the ages of the knots in the splines calculated, what penalties were applied on knot parameters, when was selectivity constant after a certain age, and when were selectivity values constrained to be zero).	The fits to the length-frequency data were poor and the Panel wished to better understand whether this related to how the spline selectivities were specified.	The knots are equally-spaced between the minimum and maximum age. Selectivity for a fishery can be set to be constant from a pre-specified age or a penalty imposed that selectivity is zero from a set age. This leads to some selectivity parameters being effectively ignored.
GG	Run the 2020 diagnostic model with the composition data omitted for the small fisheries sharing selectivity (i.e., exclude length-frequency data for fisheries 27, 29, 30, and 31).	The Panel had expected that these data would not be used when fisheries are grouped.	The run led to unexpected time-trajectories of spawning potential suggesting that the model is not very stable. However, this was caused by an error in how the catch likelihood was weighted. This is explored further in requests NN and RR.
HH	Run the 2020 diagnostic model where only selectivity for the index fisheries is assumed be asymptotic.	The 2017 assessment suggested that selectivity for the longline fisheries were not all asymptotic.	This change led to several selectivity patterns for the longline fisheries that were dome-shaped, but surprisingly to little change to the spawning potential.
II	Run the 2020 diagnostic model where selectivity for no fisheries (including index fisheries) is assumed be asymptotic.	This is a more extreme version of Request HH.	This request led to results very similar to those for request HH, but this was unexpected.
JJ	Is there a measure of the number of operations that could be used as the sample size (e.g., actual sample size, effective sample size, operations)?	The Panel wished to understand whether it would be possible to weight the length- and weight-compositions using measures that are more likely to be related to the information content of the data. The current imposition of a maximum sample size of 1,000 leads to length data for fisheries with 1,001 and 100,000+ measured fish being given the same weight	The size data are aggregated in the SPC databases so it is not possible to obtain set level resolution or even now how many sets particular sample groups come from, much of the data are port samples and not samples that can related to set number.
KK	Run the 2020 diagnostic model assigning a downweighting adjustment of 20 (not 60)	Plots of fits to the length-frequency data indicate that the 2017 diagnostic model fitted the length-frequency data better than the 2020 diagnostic model.	Very little difference in spawning potential and depletion. The fits to the length-composition data were improved, but not to the extent of the 2017 assessment.
LL	Update the plots showing the fits to the mean lengths and mean weights to include the arithmetic average mean	This will help discern patterns in these fits.	This modification to the plot was implemented and highlighted some fisheries for which the means of the observed and model-predicted lengths and weights are substantially discrepant (Figure 5).
MM	Run the 2020 diagnostic model with the Richards growth curve and a von Bertalanffy growth model with estimated deviations	This is a follow-up to some earlier requests	The model results were unrealistic with the model not mimicking the catches adequately. See Request RR.
NN	Run the 2020 diagnostic model with the grouped fisheries ungrouped	The change led to a marked change to the SSB trajectories for the 2017 model	The model results were unrealistic with the model not mimicking the catches adequately. See Request RR.
OO	Determine how many fisheries have selectivity constrained to zero.	Imposing a penalty that focuses selectivity to be zero will reduce the effective number of knots.	Several fisheries (MISC PH 7, PL JP 1, PL ALL 3, PL ALL8, MISC ID 7, PS PHID 7, and PL ALL 7), had selectivity constrained to zero
PP	Rerun the 2020 diagnostic model with the effort for regions 1, 2, 9, 5 and 6 multiplied by 2	The Panel wished to understand if the scale of the biomass in regions 1, 2, 9, 5, and 6 was driven by the assumption that catchability and selectivity for the index fisheries were the same.	The estimates of current biomass are largely insensitive to this change, while early biomass estimates are higher – the relative biomass by region matches the average CPUE (Figures F.7 and F.8)
QQ	Rerun the 2020 diagnostic model with the effort for regions 1, 2, 9, 5 and 6 multiplied by 2 and downweight the length-composition data by half	The Panel wished to understand the extent to which the scale of the biomass in regions 1, 2, 9, 5, and 6 was driven by the assumption that catchability and selectivity for the index fisheries are assumed to be same for all regions.	The estimates of current biomass are largely insensitive to this change, while early biomass estimates are higher – the relative biomass by region matches the average CPUE (Figures F.7 and F.8)

#	Request	Rationale	Response
RR	Nick to explore catch penalty and try to repeat request NN.	The results for the models for request NN led to unrealistic outcomes, suggesting that the catch penalty was not been implemented correctly.	The catch penalty was not implemented correctly (a weight of 1 instead of 100000). The results for the model that ignored the length-composition data for fisheries that are grouped and have small catches was minor. The results for the models with offsets and the Richards growth curve were similar to those of the 2020 diagnostic model but the model with offsets fit the length-composition data better, while the model based on the Richard curve fitted the conditional age-at-length data better (but the patterns observed previously remained; Figure F.9).
SS	Describe the parameterization and weights for the stock-recruitment relationship	The Panel was unclear how influential the stock-recruitment relationship was.	There is effectively no weight on the annual deviations in recruitment about mean recruitment but there is a weak penalty (CV=2.23) on the stock-recruitment relationship.
TT	How does the dynamic depletion work with different stock-recruitment curves	The Panel was unclear how (if) the stock-recruitment relationship impacted the estimates of depletion.	Dynamic depletion adjusts the recruitment estimates, but given the fairly high values for steepness, the effects are likely small.
UU	Plot the average spawning potential over time vs the average CPUE over time	The Panel wished to understand the influence of CPUE scaling on the results of the assessment.	The plot was produced (Figure 6), which shows that the average spawning potential matches the average CPUE very closely.

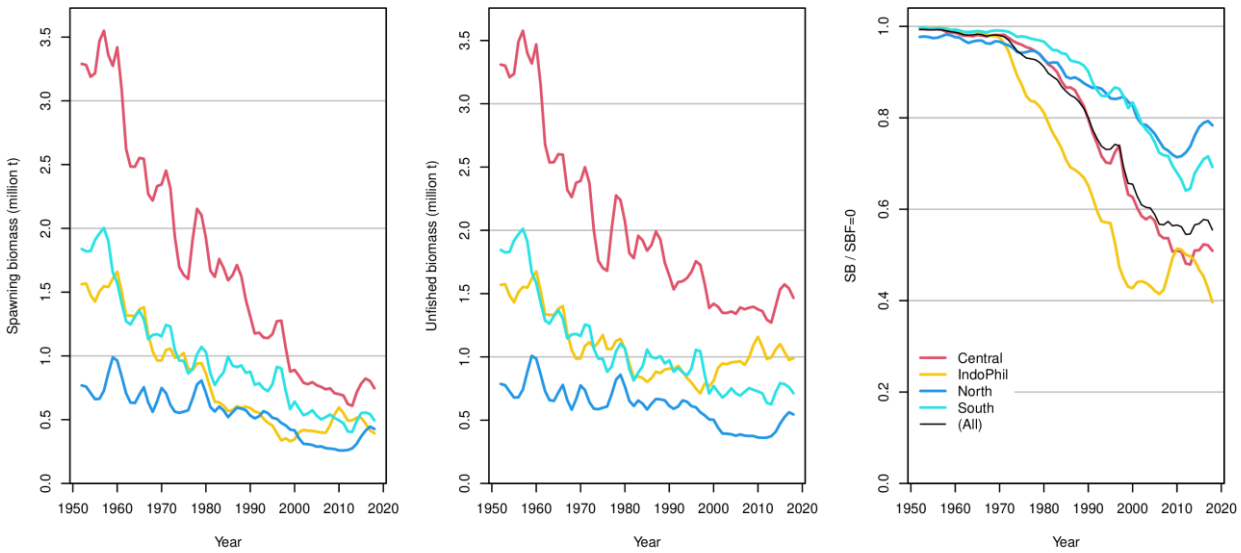


Figure F.1. Spawning potential, unfished spawning potential and depletion for the north, equatorial (central), Indonesia/Philippines (IndoPhil) and south areas. Central = model regions 3, 4, 8; IndoPhil = model region 7; North = model regions 1, 2; South = model regions 5, 6, 9.

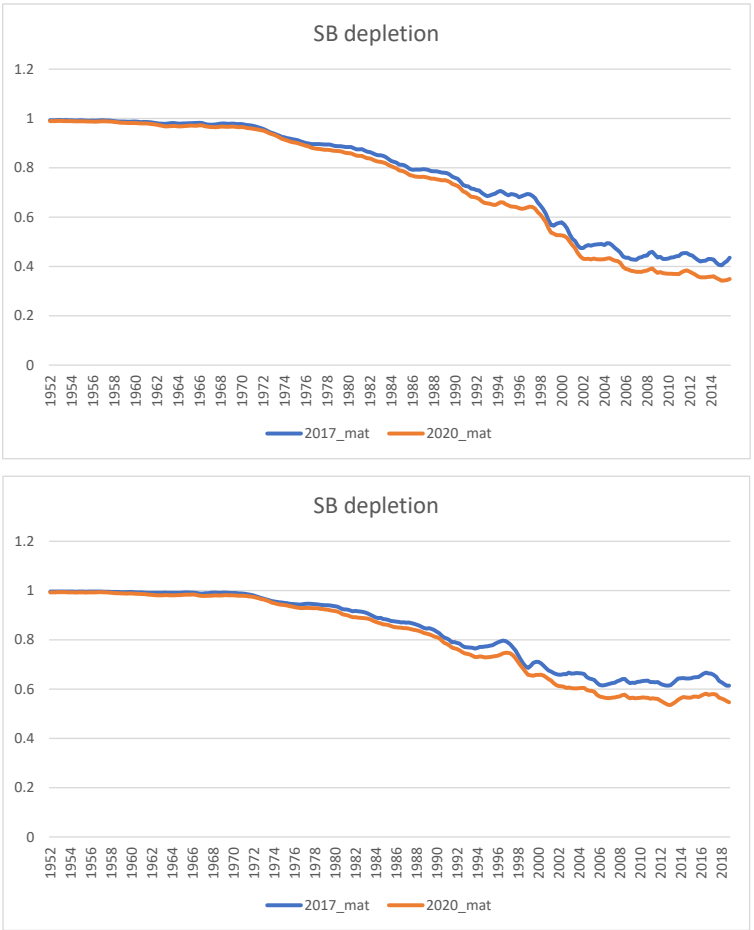


Figure F.2. The results of the 2017 diagnostic model (2017_mat) and when its results are based on the 2020 maturity-at-age vector (top figure), and the results of the 2020 diagnostic model (2020_mat) and when its results are based on the 2017 maturity-at-age vector (2017_mat; bottom figure).

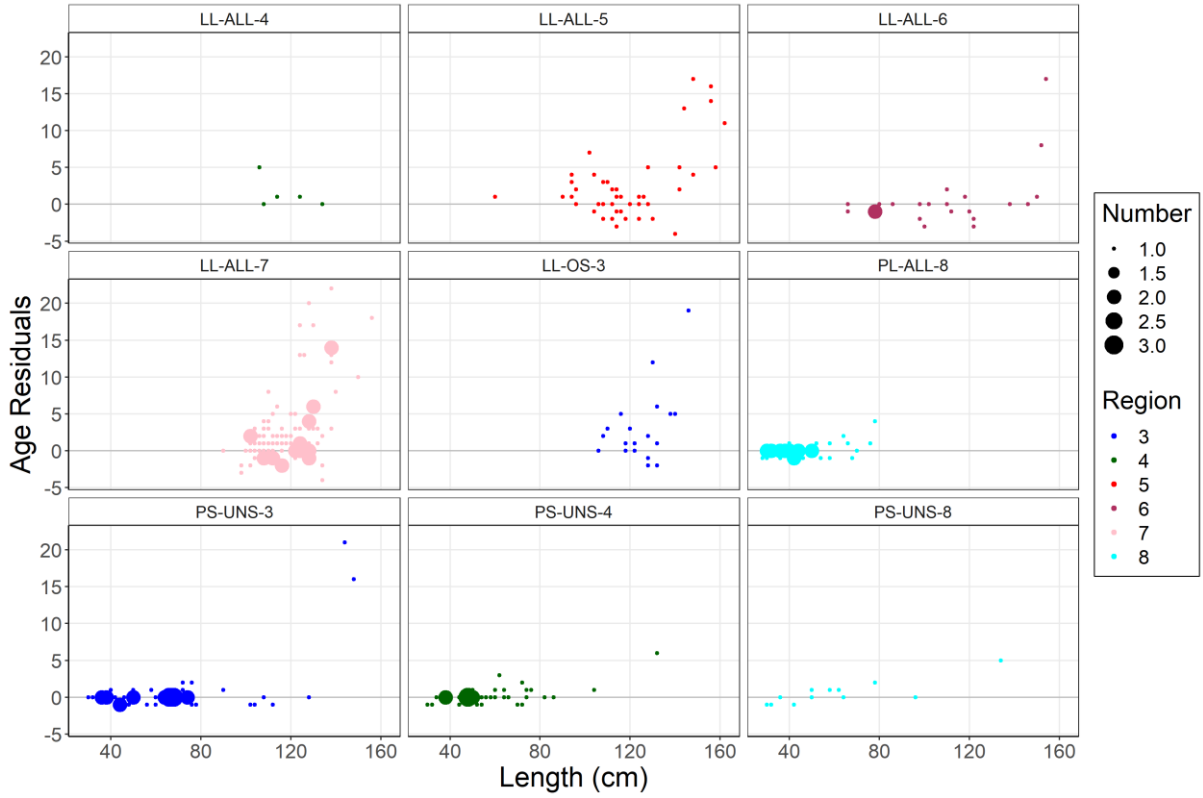


Figure F.3a. Conditional length-at-age residuals for the 2020 diagnostic model (coloured, sized by number).

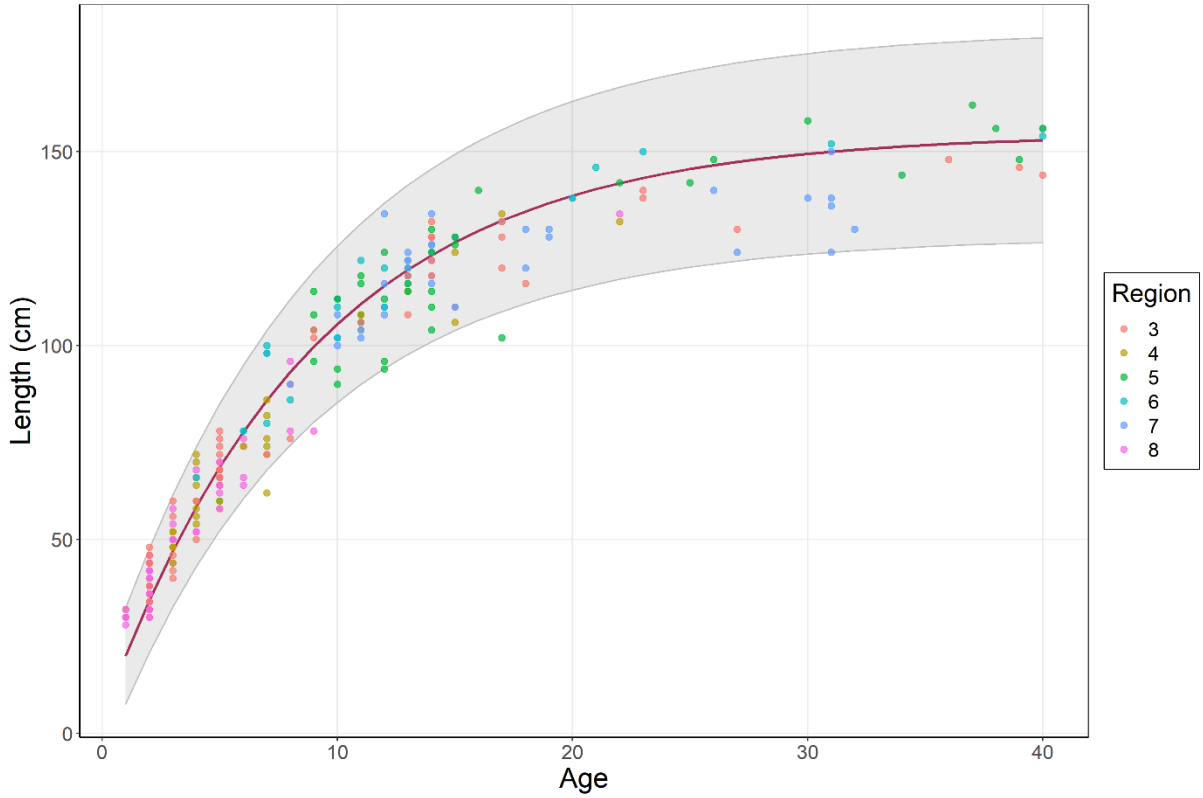


Figure F.3b. Growth curve with 95% confidence interval for the 2020 diagnostic model.

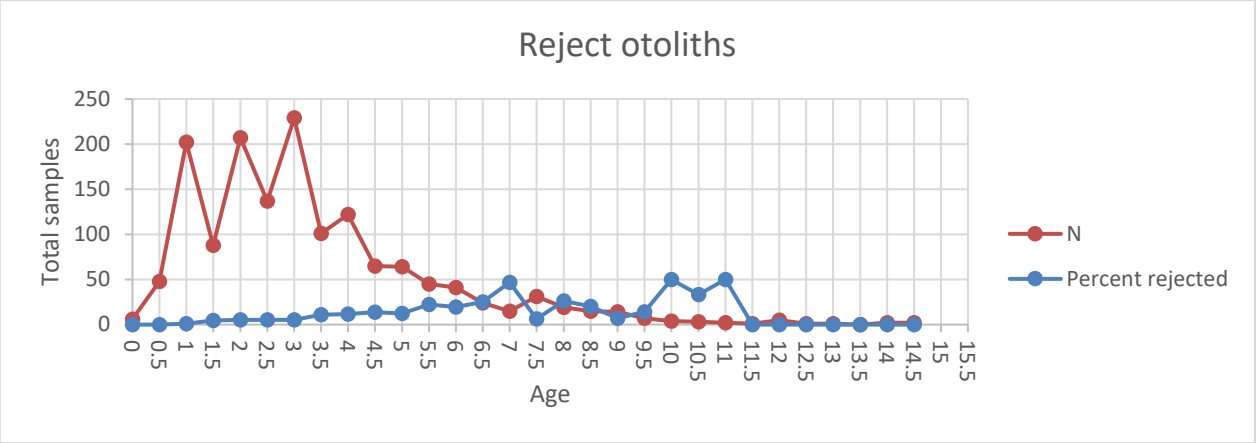


Figure F.4. Number of animals sampled for aging and the proportion rejected because of a lack to reach agreement on age.

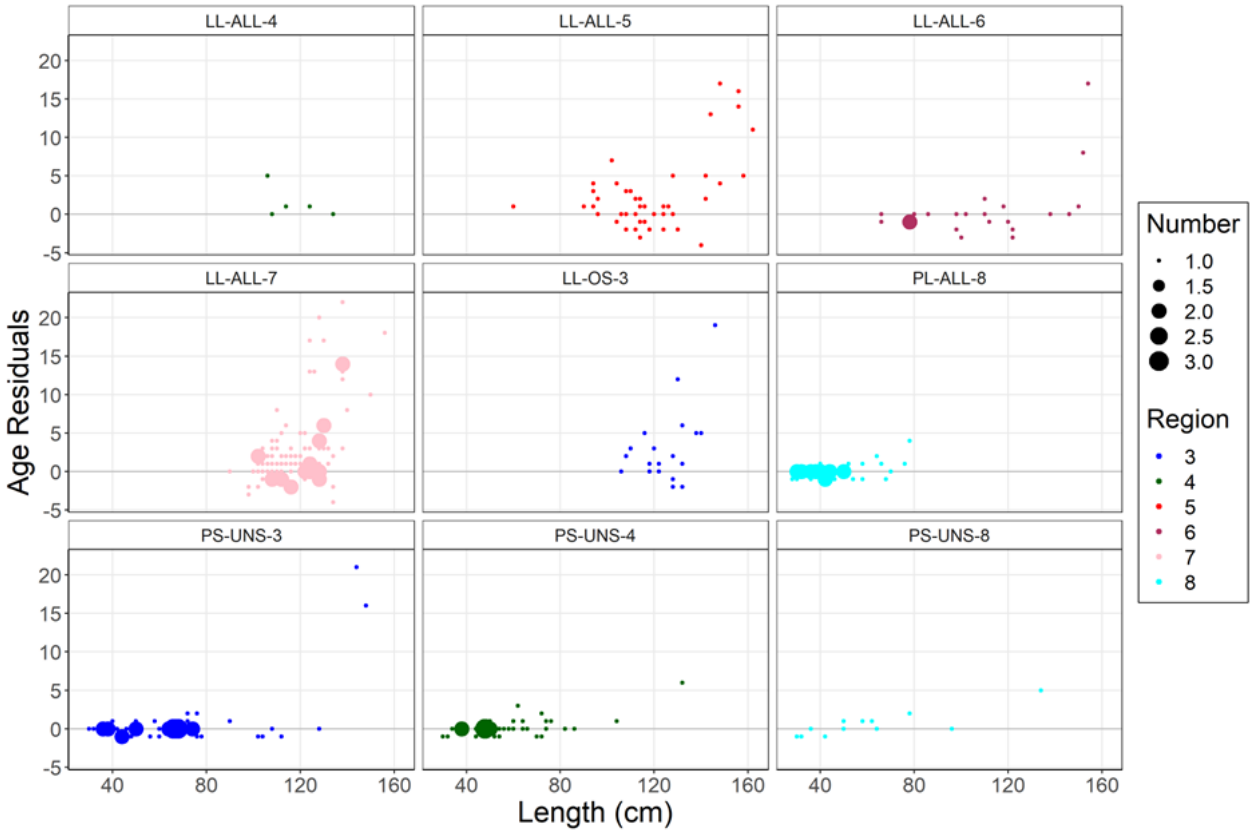


Figure F.5. Conditional age-at-length residuals for the 2020 diagnostic model (colored by region, size proportion to number of samples).

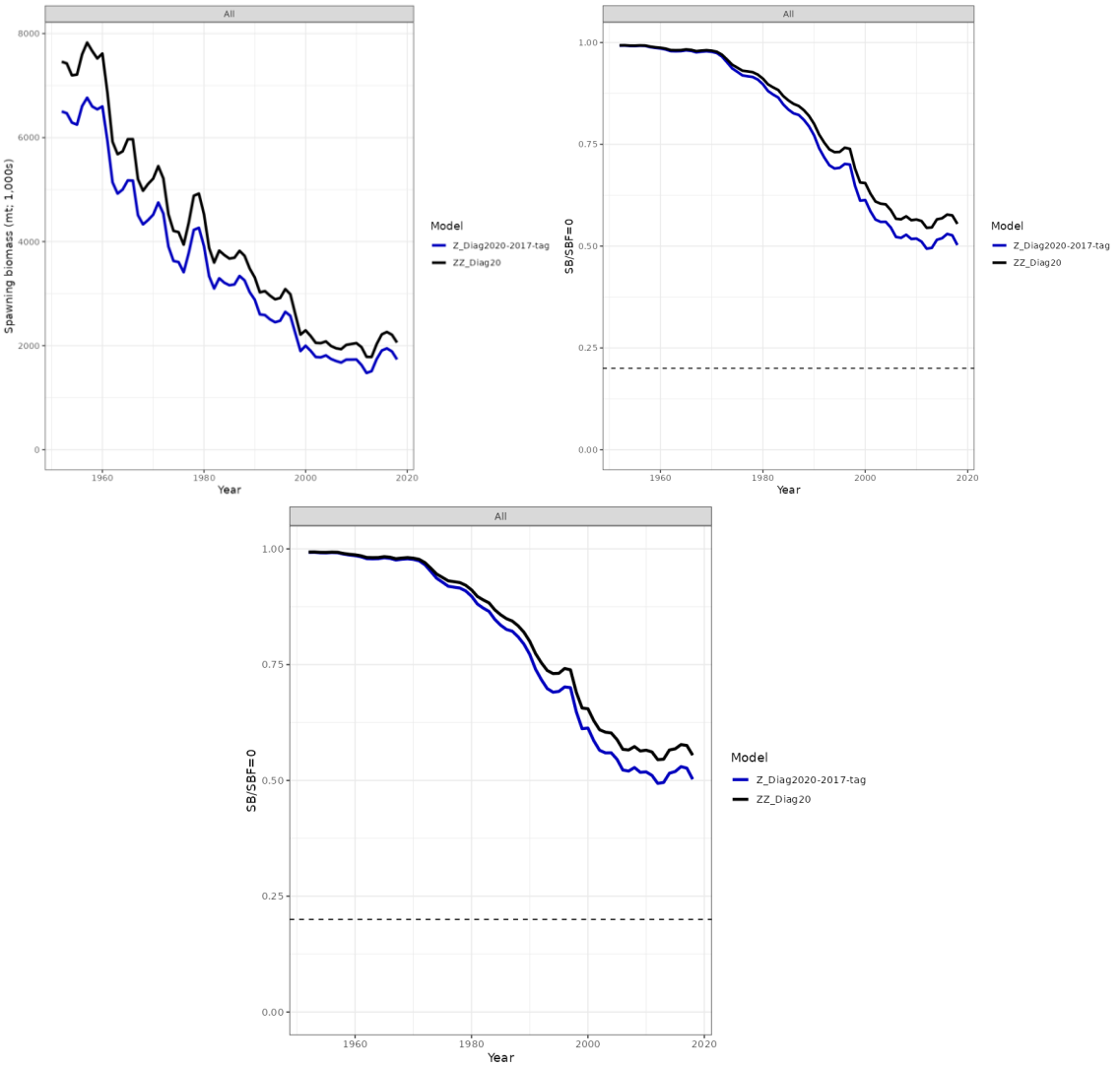


Figure F.6. Spawning potential, unfished spawning potential and depletion when the 2020 diagnostic model is based on the 2017 tagging data.

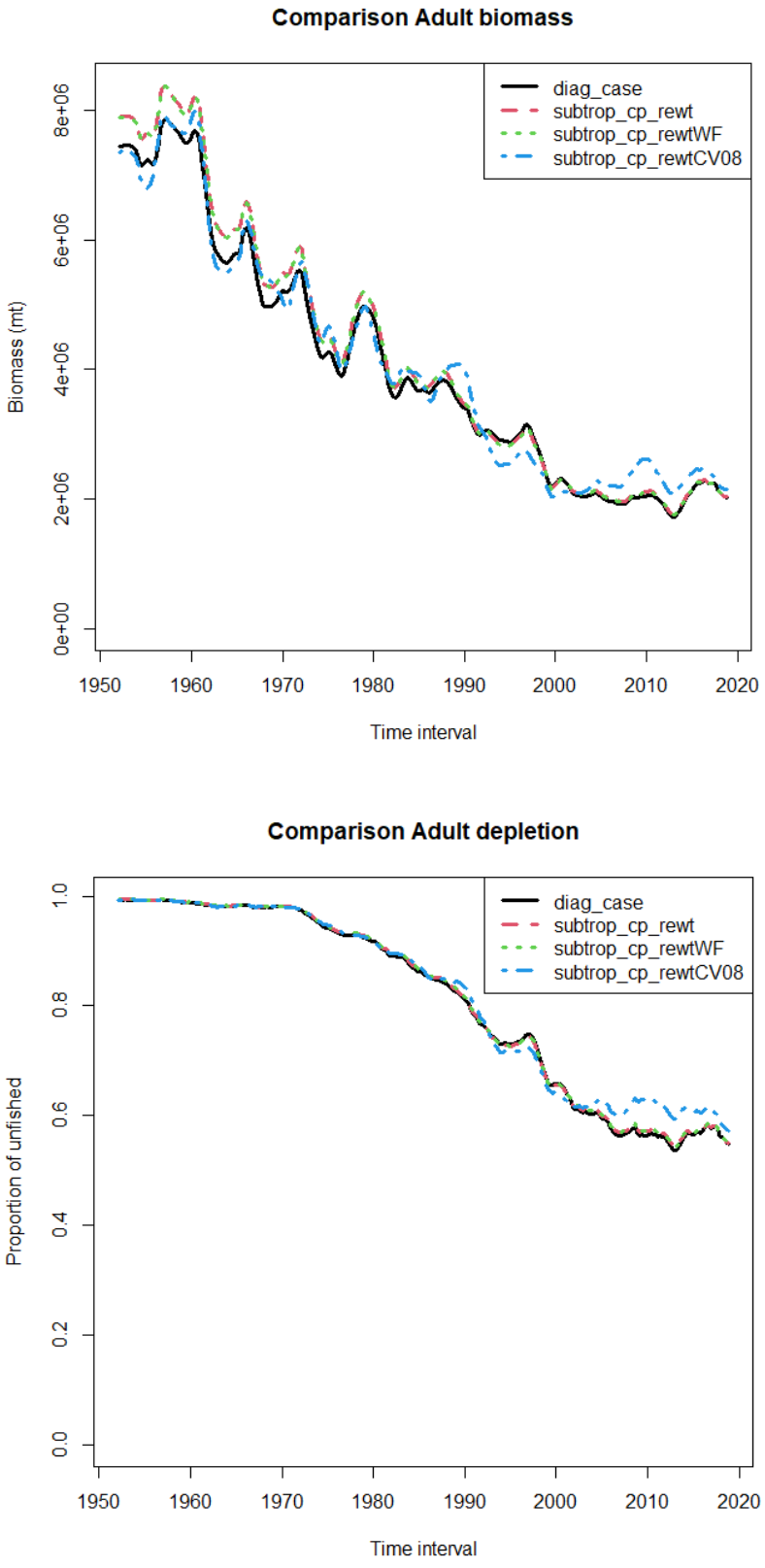


Figure F.7. Comparison among models for the total adult biomass trajectories (top panel), and the trajectories of adult biomass depletion (bottom panel) for the 2020 diagnostic model and the models of requests PP and QQ, and a variant of the request QQ model with lower weight on the CPUE data.

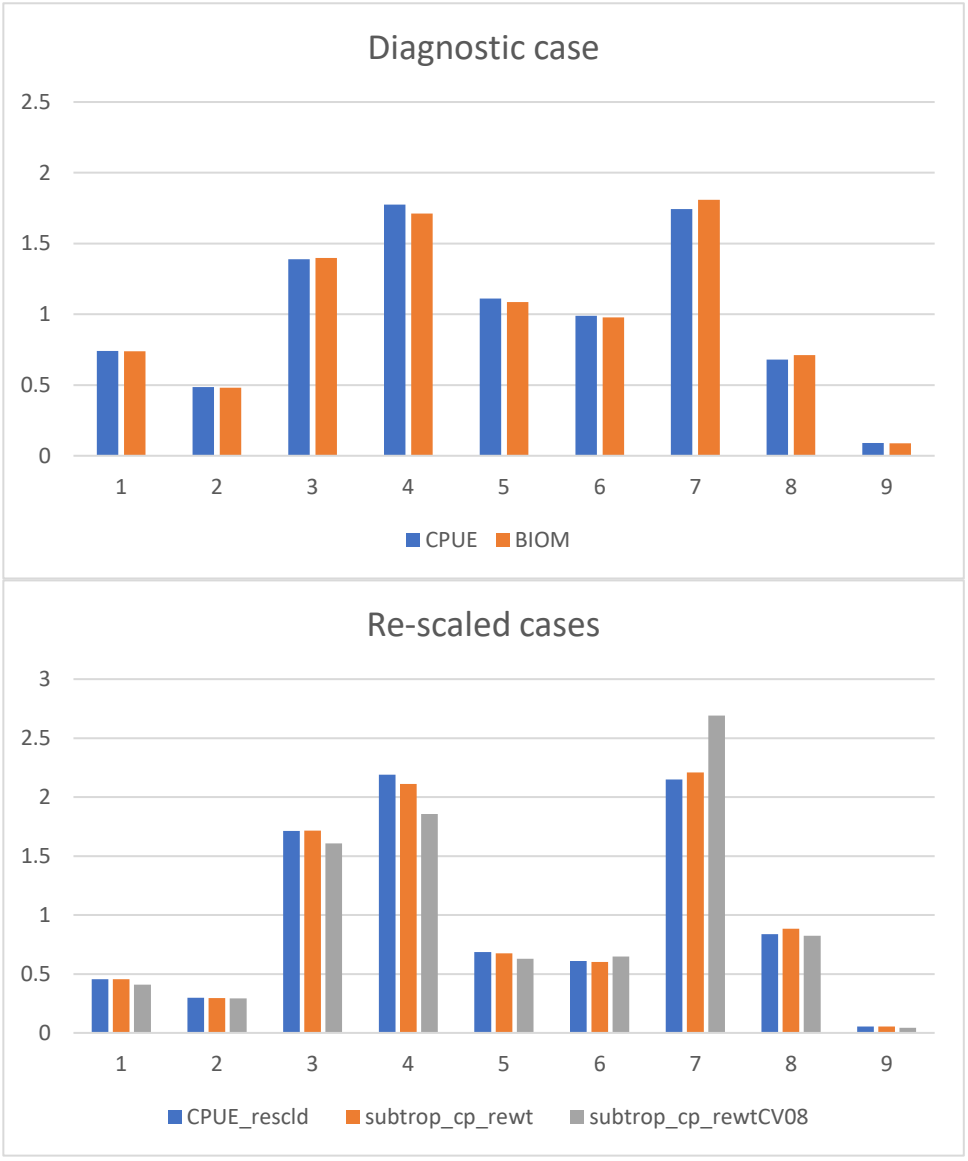


Figure F.8. Comparison among models for the normalised mean CPUE and regional adult spawning potential for the 2020 diagnostic model and the models of requests PP and QQ.

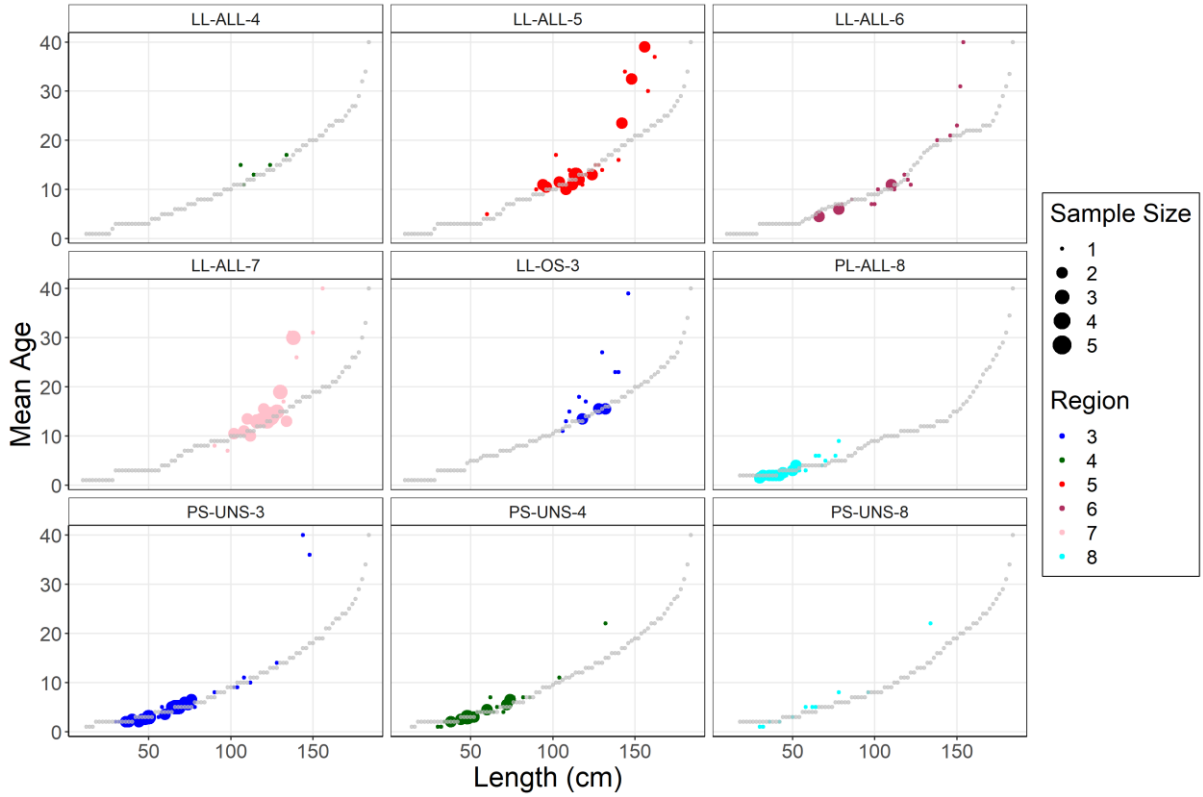


Figure F.9. Conditional age-at-length residuals for the variant of the 2020 diagnostic model that included the Richards growth curve (colored by region, size proportion to number of samples).

Appendix G. Summary of comparison of stepwise data inputs

The stepwise model developments changes introduced in the 2020 assessment that had the greatest effects are detailed below:

09_IdxNoeff … 10_SelUngroup

Flags

87 flags were changed, related to selectivity grouping, selectivity shape, length sample size, and catchability deviations.

Model setting	Applies to	Before	After	Flag
Selectivity grouping	Fisheries 4, 9, 11, 12, 29	Grouped	Split into four groups	Fish flag 24
”	Fisheries 13, 15, 24, 25, 30	Grouped	Split into four groups	Fish flag 24
”	Fisheries 17, 23, 32	Grouped	Split into three groups	Fish flag 24
”	Fisheries 10, 27	Ungrouped	Grouped	Fish flag 24
Selectivity shape	Fishery 7	Non-decreasing with age	Can decrease with age	Fish flag 16
”	Fishery 28	Zero for all ages over 24 quarters	Not constrained to be zero	Fish flag 16
Length sample size	Fisheries 7, 8, 29	Divisor = 20	Divisor = 40	Fish flag 49
Catchability deviations	Fisheries 1-41	Constant for 24 months after each change	Can vary between quarters	Fish flag 23

Data

(No changes to data.)

10_SelUngroup … 11_JPTP

Flags

606 flags were changed, related to selectivity shape and adding tag groups.

Model setting	Applies to	Before	After	Flag
Selectivity shape	Fisheries 1, 5, 6, 9, 10, 12, 27	Can decrease with age	Non-decreasing with age	Fish flag 16
”	Fisheries 17, 20, 21, 22, 23, 24, 28, 32	Not constrained to be zero	Zero for all ages over 25 quarters	Fish flag 16
”	Fisheries 2, 4, 5, 6, 7, 9, 11, 12, 29	Not constrained to be zero	Zero for ages 1-2 quarters	Fish flag 75

Tag file

- Define mixing period of the tag recaptures to be 182 days for each tag release.
- Tag release groups are increased from 87 to 145 because JPTP program tags were added.

- There are more recaptures for RTTP and PTTP programs. This is a change in the 2020 assessment, including tags without recapture locations in the purse seine fisheries, as well as tags added after revising the tagger effect analysis.
- Number of effective releases are higher for some length bins in all programs. This may be a change to usability correction for having additional recaptures.

Ini file

- With more release groups, the ini file has more lines to assign reporting rates, priors, and penalties for those additional release groups.

Frq file

- Number of release groups was updated.

11_JPTP … 12_Age10LW

Flags

41 flags were changed, related to selectivity shape.

Model setting	Applies to	Before	After	Flag
Selectivity shape	Fisheries 1-41	Constant for all ages over 25 quarters	Constant for all ages over 37 quarters	Fish flag 3

Ini file

- Number of age classes increased from 28 to 40 quarters.
- Maturity-at-age updated for all age classes and extended for the increased number of age classes.
- Natural mortality is slightly decreased from 0.25 to 0.23.
- Age parameters updated for all age classes and extended for the increased number of age classes.
- Length-weight parameters updated.

12_Age10LW … 13_CondAge

Flags

196 flags were changed, related to growth curve estimation, initial population, fishing mortality, catch likelihood, and selectivity shape.

Model setting	Applies to	Before	After	Flag
Growth curve estimation	Parameters	Not estimated, apart from variance	Not estimated	Parest flag 32
”	Early ages	First 8 quarters are independent parameters	All ages follow growth curve	Parest flag 173
”	Penalty	No penalty wt for length estimation	Penalty wt of 1 for length estimation	Parest flag 182
”	Age-length data	Model fit to observed data not activated	Model fit to observed data activated	Parest flag 240
Initial population	Scaling pop	Estimated	Disabled	Age flag 113

Model setting	Applies to	Before	After	Flag
Fishing mortality	Max F	0.7	5.0	Age flag 116
Catch likelihood	Common wt	100,000	10,000	Age flag 144
”	Specific wt	0	100,000	Fish flag 45
Selectivity shape	Fisheries 17, 23-24, 28, 32	Constant for all ages over 37 quarters	Constant for all ages over 12 quarters	Fish flag 3
”	Fisheries 20-22	Constant for all ages over 37 quarters	Constant for all ages over 24 quarters	Fish flag 3
”	Fishery 6	Logistic shape	Cubic spline, or length-specific selectivity	Fish flag 57
”	Fishery 28	4 spline nodes	5 spline nodes	Fish flag 61

Age-length file

- Addition of otolith data through conditional age-at-length.

Ini file

- Updated maturity-at-age.
- Natural mortality updated.
- Age parameters updated.
- Length-weight parameters updated.

13_CondAge … 14_MatLength

Flags

1 flag was changed, related to maturity.

Model setting	Applies to	Before	After	Flag
Maturity	Convert mat @ length to age	Not converted	Converted using weighted spline	Age flag 188

Data

(No changes to data.)

14_MatLength … 15_NoSpnFrac

Flags

(No flags were changed in this step.)

Ini file

- Updated maturity at length.

15_NoSpnFrac … 16_Size60

Flags

83 flags were changed, related to length sample size and weight sample size.

Model setting	Applies to	Before	After	Flag
Length sample size	Fisheries 1-2, 4, 7-9, 11-12, 29, 33-41	Divisor = 40	Divisor = 120	Fish flag 49
Weight sample size	Fisheries 1-2, 4, 7-9, 11-12, 29, 33-41	Divisor = 40	Divisor = 120	Fish flag 50
Length sample size	Fisheries 3, 5-6, 10, 13-28, 30-32	Divisor = 20	Divisor = 60	Fish flag 49
Weight sample size	Fisheries 3, 5-6, 10, 13-28, 30-32	Divisor = 20	Divisor = 60	Fish flag 50

Ini file

- Updated tag reporting group flags.
- Updated maturity at age and maturity and length.
- Updated von Bertalanffy parameters.

16_Size60 … 17_Diag20

Flags

12 flags were changed, related to selectivity shape.

Model setting	Applies to	Before	After	Flag
Selectivity shape	Fisheries 13-16, 19-22, 25-26, 30-31	Not constrained to be zero	Zero for age 1 quarter	Fish flag 75

Data

(No changes to data.)

Appendix H. Summary of panel key recommendations and suggestions and SPC responses.

Panel key recommendations and suggestions	SPC response
Resourcing: The Panel consequently recommends that the analysts be given more time or additional technical support to ensure that model exploration is such that it is possible to fully understand the causes of changes in model results.	More time would be valuable given the amount of work involved with these large tuna assessments and finalization of data sets in April, while the assessment report is expected in July. However, it seems there is little scope to alter the SC schedule, and finalising data sets earlier in the year seems problematic for the moment. One option is to do as much of the stepwise (bridging analyses) on the previous assessment data, then apply data updates when the new data is finalized. We will do this for yellowfin and bigeye in 2023. Provision of additional funds to support assessment work would be useful, particularly to ensure staff resources are sufficient to conduct follow-up work from previous assessments and conduct data preparation and analyses for the assessments. This would relieve pressure from the lead assessment scientist and allow more focus on the modelling challenges. For 2023 we will also move the SPC pre assessment workshop a month later (25-27 April). This will hopefully allow the assessment scientists to have more material to present and discuss. We note the SC18 has also requested that we provide a paper on options for allowing more time for SC participants to review assessments prior to SC meetings.
Model changes: A step in a bridging analysis should involve a single change only, with perhaps "minor" changes to the MULTIFAN-CL settings that are needed to ensure convergence. Table 1 provides a simplified example format that may aid in documenting changes to model specifications. The Panel recommends a table like this (or something similar) be adopted for future assessments so that effects can easily be understood and isolated.	We agree with the panel recommendation and will endeavor to improve the documentation and approach to the bridging analyses. This should also include information on changes to data inputs and analysis methods.
Natural mortality: The Panel recommends continuing the current approach with the base *M* for the Hoyle et al. (2009) method set to 0.2 quarter^{-1} but including alternative values for base *M* in the uncertainty grid. The range of base *M* values could be determined using a likelihood profile or the bounds from Hoyle et al. (2022), but for now there is no basis to set this value other than to the default of 0.2 quarter^{-1}.	We concur with this recommendation. We support increased efforts to collect high quality data on sex ratio at length, as this data is critical for fitting *M* at age, and existing sex ratio data is limited, especially for larger fish. The review paper from Hoyle et al. will be very useful.
Growth: The sampling process for age data aimed to obtain a similar number of otoliths per length-class, which means that an externally fitted growth curve will be biased. The Panel therefore recommends not basing the growth curve on an external estimate unless internal estimates are clearly implausible or an appropriate sampling approach to obtain representative population length-at-age data can be developed and implemented, or a growth curve is externally estimated using conditional age-at-length data.	We note this concern regarding selective population sampling for otoliths, and that the recommended approach is to use conditional age-at-length data to inform estimates of growth, when otolith readings are sufficiently reliable. We expect to estimate growth using conditional-age-at-length data in the 2023 assessments for the diagnostic case but note there is still ongoing otolith age validation work. We also note this approach is not possible for skipjack tuna currently because reliable otolith-based conditional age-at-length data are not yet available. Developing an appropriate sampling design and investigating the impacts of the quantity and spatio-temporal coverage of the conditional age-at-length data is warranted. Sampling would

	benefit from greater contribution across the WCPFC membership to collect samples and provide to the SPC tissue bank.
Size composition: No specific recommendations were provided but the Panel discussed the issue of the weighting scheme for composition data. They noted "*The current weighting scheme gives equal weight to all composition data sets with a sample size (number of fish measured) of at least 1,000. Moreover, the weighting scheme does not account for how the samples were collected (e.g., large numbers from a few sets/trips or small numbers from many sets/trips). Consideration should be given to weighting the composition data using a metric that reflects the likely information content of the data (such as sets/trips), but this will require access to more basic data than is currently available to the analysts."* They also noted that, *"Therefore, the sample size of the composition data should be analyzed outside the stock assessment model (e.g., using bootstrap analysis or spatio-temporal models) or the appropriate measure of sample size chosen (e.g., number of sets or trips"* and *"These approaches generally assume that the relative among-year sample size is maintained, but modifications to the maximum have been proposed (e.g., asymptotic functions). The best approach has yet to be determined and is still an active topic of research ".*	We have begun exploring options for alternative measures for input (stage 1) sample sizes for size composition data. However, the current data collection protocols across such a large fishery and diverse fleets, with both observer and port sampling present some problems. It may be possible to use number of trips for long line size data. For purse seine fleets, most data is aggregated across multiple purse seine trips for space/time strata (i.e. quarter/fishery/region) due to the grab sample approach where very few fish are measured from individual sets, and if port sampling is conducted the number of trips might be feasible, but port sampling has not been consistent across time and fleets.. But we agree there is a need to consider alternative approaches for assigning appropriate input sample sizes that better reflect the relative information content in the composition samples.
CPUE: The panel noted that the model does not fit the mean weights for the index fisheries for regions 3, 4 and particularly 8 well. The reasons for these discrepancies are unclear but may be related to the assumption that selectivity and catchability are assumed to be the same for all regions. Some relaxation of this may be necessary to resolve this (see Section B.1 below). The panel, while supporting the continued use of VAST to model CPUE, had concerns about how poorly sampled cells are included in the analysis. The panel made the following suggestions: Run the spatio-temporal model (e.g., VAST) by region, and (a) compare correlation between the regionally estimated indices (independently) with the same regions split up from global model results, (b) compare decorrelation distances among regions and see how different they are from the global estimate, and (c) assess the extent that within-region trends differ from the global trends. Examine the extent to which the current indices are correlated owing to their being computed from one model and reflect this (if substantial) by a variance-covariance matrix when fitting to the data. Examine if covariates can be categorized by abundance and catchability, (b) determine how covariates affect the model, (c) consider including interaction terms, and (d) include a quarterly random effect, perhaps in a hierarchical approach Consider running the spatio-temporal model within a (main) region for all fleets and compare the results to those from a run with only a principal fleet included to assess the effects of combining data for multiple fleets into a single analysis. Further evaluate both the definition of viable cells and how VAST shares information for cells and times with little information. This is particularly important for evaluating the size of the north and south regions and the influence of edge effects in the CPUE standardisation.	We will initially focus on increasing confidence that the VAST abundance indices developed from the longline CPUE data do a good job at representing spatial differences in relative abundance among model regions. This will involve analyses that deal with most of the suggestions listed. In terms of trends and dynamics of the indices, exploring covariates for catchability and availability will continue as part of the CPUE analyses. Environmental variables can be further explored as well as gear covariates. These explorations may initially involve simpler exploratory analyses to detect relationships before building more complex models and will depend on the availability of data on gear and environmental covariates throughout the time series. Environmental data such as SST and DO (dissolved oxygen), or even predictions from the SEAPODYM model, may also be used to define unsuitable yellowfin (or bigeye) habitat as a guide to exclusion of geographic areas prior to fitting the VAST models. In relation to the suggestion on relaxing the assumptions of constant selectivity and/or catchability across longline fleets in the CPUE analyses, we will need to consider this further before such an approach is adopted. Outcomes of the CPUE modelling investigations may also result in improved confidence in the abundance indices and the assumptions required to use these as information on relative regional biomass in the model.

Effort creep: The Panel suggests that sensitivity analyses should be conducted to explore what levels of effort creep are required to influence management quantities, but that the effort creep scenarios applied in models used for management advice should have a sound basis. The Panel therefore recommends that the SPC is supported to conduct further investigation of effort creep in the longline and purse seine fisheries. This will require support from Distant Water Fishing Nations for catch and effort data provision and information on how operations, vessel features, gear and technology uptake has changed over time for their fleets. The Panel understands a proposal will be submitted from SC18 to the WCPFC19 for support to study effort creep. This study is based on recommendations from the 2022 WCPO skipjack assessment and would focus on pole and line and purse seine fisheries. The Panel recommends that this project be expanded with additional funding to also consider longline fisheries.	We will further consider how effort creep can be accounted for in the CPUE standardisation for the longline fisheries, but we expect work to document, understand and develop well founded effort creep scenarios will need to continue beyond the upcoming assessments. Sensitivity analyses of effort creep scenarios for longline CPUE could be included in the upcoming assessment, informed by discussion at the Pre-Assessment Workshop. SPC does not have direct access to the major fleets comprising the longline fisheries so detailed analyses of effort creep will require collaboration with Distant Water Fishing Nations (DWFN). We suggest that a dedicated project on longline effort creep should be proposed by the SC19 involving DWFNs and we would encourage interested scientists to contribute to this work.
Tagging data: The panel focused, in particular, on issues around reporting rates and tag mixing periods. The panel agreed the approach of using the number of days at liberty, rather than the number of quarter boundaries crossed, is more reasonable as a basis for defining tag mixing periods. They questioned the basis for the choice of any particular fixed mixing period assumption and suggested the approach to allocating variable mixing periods based on individual-based modelling (IBM) as done for skipjack tuna by Scutt Phillips et al. (2022) provides a more defensible basis for assigning mixing periods, rather than relying on fixed assumptions for all tag releases. The panel noted that there is a need for more tag-seeding experiments, to help with tag reporting rates that were estimated on their upper bounds for several regions/fisheries, and that tag seeding efforts should be prioritized according to where catches are most important. The panel suggested that fisheries with multiple tag-reporting rates over time could be treated as multiple fisheries, each with a time-invariant tag-reporting rate or MULTIFAN-CL could be modified to allow for time-variation in the tag-reporting rates.	We will continue recent SPC practice to use the number of days at liberty to define mixing periods. How to estimate the times at liberty after which tagged fish can be considered fully mixed with the untagged population continues to be challenging. The advances in this area for the 2022 skipjack assessment are an improvement, but the SEAPODYM model for yellowfin is not sufficiently developed to support this approach for the next assessment. While the bigeye SEAPODYM model is more advanced than yellowfin, the issue of tag mixing is of less importance/influence for the bigeye assessment, due to the limited tagging data, and we do not think the amount of work is justified to develop the skipjack approach for bigeye. At this stage it seems likely that a range of fixed mixing periods will be applied in the next bigeye and yellowfin assessments, consistent with the 2020 assessments. We will review how we assign reporting rates for fisheries, but it is not possible to modify MFCL to allow for time variation in reporting rates in time for the 2023 assessments.
Catches: There is uncertainty in the Philippines and Indonesian catches prior to 1990 and investigation of approaches to improve these estimates, or include the uncertainty in the assessment, should be continued, perhaps through the WPEA (West Pacific East Asia) project. Other tRFMOs receive longline catch in weight or numbers in different years and it should be confirmed that the data received by SPC is received in the units that catches were measured in and not pre-converted by the member states.	We acknowledge there is uncertainty in catch data for the Indonesia/Philippines/Vietnam areas. While work with these countries under the New Zealand MFAT funded WPEA (Western Pacific East Asia) data improvement project is improving the situation, notably for the Philippines, this is challenging considering the uncertain and limited resources for these countries to collect data and more work is needed. As to historical uncertainty, we concur this should be considered. Noting that we currently have limited understanding of the level of uncertainty or bias in catch estimates, it is possible to explore the sensitivity of the model estimations to catch uncertainty, noting this adds additional time to the assessment and would not necessarily feature in management advice. This may be discussed at PAW. We will continue to support work on providing estimation of historical and current uncertainty in the reported catches through the

	WPEA project. We will confirm through our data team, and if necessary DWFNs whether data received by SPC is received in the units that catches were measured by the member states.
Selectivity parameterization: The panel was concerned with poor fits to aggregated size composition data. The Panel recommended the following approaches: a. Define fisheries using a regression tree analysis applied to the composition data (e.g., Lennert-Cody et al. 2010, 2013; Maunder et al., 2022). b. Describe the fisheries, including the magnitude of catch, sample size, and whether it is an index. c. Triage the composition data to remove data that are likely to be unrepresentative and/or unreliable. This may include excluding data for a whole fishery (and sharing selectivity), entire years, or for some lengths (e.g., when small amounts of fish under the minimum legal size are caught). d. Avoid aggregated compositions that show multiple modes, shoulders, or other unusual patterns (i.e., not logistic or double normal) by separating them into more fisheries or allowing for time-varying selectivity. e. Assume that selectivity is length-based unless it is known to be age-based (e.g., due to ontogenetic movement). f. Ensure that the composition data for fisheries that catch a large proportion of the catch are fit well. This might require the use of flexible time-varying selectivity. g. Consider downweighting the composition data for fisheries with low catch as these do not need to be fit well. h. Fit the composition data for indices well - consideration should be given to allowing selectivity to be more flexible and time-varying if necessary. i. Use the empirical selectivity diagnostic (Maunder et al., 2020) to check that selectivities are appropriate. j. Use the empirical selectivity method to determine the number of knots and their position when using splines. k. Consider dropping the composition data for fisheries whose selectivities are shared with those for another fishery because the data for those fisheries are considered inadequate, particularly if it has low catch. The panel also commented that the assumption of shared selectivity and catchability among index fisheries is problematic because there is evidence of differences in growth among regions, which might be best modelled using different selectivity patterns for each region. However, this would mean losing the information on relative regional scaling. Consideration should be given to allowing for some differences among regions while maintaining similarities to retain information on regional scaling. For example, catchability and selectivity for each region could be modelled as a penalized deviate from the overall mean or one region set as the reference for catchability and selectivity and the other regions deviating from that region. Selectivity might require age-/length-specific deviates or something more complicated with either the peak changing or a functional form describing offsets for all ages/sizes. It is unknown if this is possible in MULTIFAN-CL.	The poor fits to the aggregated length compositions are a key concern for the panel. This requires close attention for the next assessment. Poor fits to size composition can lead to biased model outcomes. The review panel provides many suggestions, we respond to each below. a) We will investigate this analysis b) This information is included in the supporting papers for the assessment, additional data on catch magnitude can be included in the fishery definitions table in future assessments c) We are reviewing the composition data, both to better understand the representativeness of the spatio-temporal sampling. This may result in filtering out some size composition data. There are no minimum legal length limits for tunas in the fisheries relevant to the assessments, and we think that excluding data from specific length categories is not warranted. d) Areas a) and c) may help with removing some of the unusual patterns in the size distributions, particularly earlier years. e) Length based selectivity seems a preferred option but requires development work on MFCL. We are investigating the feasibility of this option. f) Time varying selectivity is possible through greater use of time blocks. It is also possible by estimating selectivity deviations, but involves a large number of parameters and seems not to work with the catch-conditioned approach at the moment. g) This will be explored. h) As per f) i) We are not familiar with the empirical selectivity diagnostic referred to that apparently has been developed for SS3. We will look into this. We expect we will focus this assessment on triaging the size composition data, reviewing the fishery definitions and the need for time blocks for selectivity, and implementing size-based selectivity. j) As for i) k) This is a good suggestion that we will consider when reviewing the size composition data and fishery definitions. There are likely to be some poorly sampled

	fisheries, with low catches with shared selectivity, that are difficult to fit. The issue of shared selectivity for the index fisheries is problematic. The best approach to dealing with a relaxation of this assumption while still allowing the CPUE to provide information on regional scaling of population sizes will require further consideration using MFCL.
Recruitment: The assessment model fixes the recruitment for the recent six quarters to the mean. This may influence the results if there is information about recruitment in the data. The model should be run estimating these recruitments to determine the impact on the results. The panel provided a commentary and the appropriate use of the log-normal bias correction. They also discussed the alternative approaches to treating early recruitments where information is limited – either setting recruitment deviates for early years to zero or estimating them, where the latter my then compensate for model misspecification. They note that, this approach is also associated with the selection of the start time of the model and the method used to create the initial age-structure. The best approach has yet to be determined and requires further research.	We can run models with recruitment estimated until the terminal time interval as a sensitivity. In relation to early recruitment deviations, we can also explore sensitivities where these are set to zero. However, there are diagnostics that can be looked at to help determine when to stop estimating recruitment - and we would prefer that approach. The log normal bias correction for recruitment predictions is available in MFCL and is applied as a standard practice, including in the yellowfin tuna assessment.
Movement: The panel commented that movement likely differs between adults and juveniles. Future work should look at releases and recaptures by size groups to identify any differences. Age-specific movement should be investigated in the assessment either by fixing movement to zero for adults or estimating age-specific movement.	The 2020 yellowfin assessment did not include age specific movement. This could be explored in the 2023 assessment, such as setting adult movement to zero. Other scenarios might be suggested through the conceptual model explorations.
Hessian matrices: The panel notes that the computation of the Hessian was missing, along with the analogous approximate asymptotic variance (and covariance) estimates. Several suggestions were provided on how this might be improved (e.g., modifying the configurations so that parameters were not on the bounds, trying a generalized inverse for the Hessian to obtain correlation estimates).	This will be an important area to improve in the next assessment. Implementation of a catch conditioned model as applied in the 2022 skipjack assessment will reduce the number of parameters to be estimated which will hopefully assist in achieving a positive definite Hessian. Other improvements such as ensuring tag group reporting rates are not estimated close to bounds appear promising.
Model complexity: The panel notes that model complexity is one of the main reasons the WCPFC and SPC requested an external review. The panel noted the model structure for yellowfin mimics that of bigeye largely for reasons of efficiency but the structure of the data for, and the behavior of, yellowfin differ from those of bigeye, such that the current model structure for yellowfin likely leads to model instability and unnecessary complexity. The panel made several recommendations on model complexity (page 10-11). The key aspects they recommended in abbreviated form were: a. Develop a conceptual model for yellowfin tuna in the WCPO. b. Analyze composition data to assess which areas/fleets should be combined for the purposes of defining regions and fisheries based on methods of Maunder et al. (2022) and Lennert-Cody et al., (2010, 2013). c. Simpler model structure may result from a) and b) and there seems little basis for separating region 9 for this assessment. d. Use the conceptual model to identify some realism constraints and expected model behaviour, i.e. regions where recruitment should be expected.	We will consider options for reducing model complexity that are compatible with the biology and data availability. The conceptual model will be a useful exercise and provide a background to the choice of spatial structure hypotheses, while also allowing recognition of the level of understanding of biological population structure and dispersal/movement behaviour with age. Since the previous assessment we have conducted an analysis of spatial CPUE patterns using Convergent Cross Mapping and will attempt to complement this with the analyses of length composition data following the Lennert-Cody et al. method. These analyses will be considered in developing a conceptual model for yellowfin population and fishery structure and may lead to proposed simplified model spatial structure. We don't expect to have the time available in 2023 prior to the SC19 to run models in different platforms such as Stock Synthesis but do intend to explore simpler spatial structures. Realism constraints are useful, and we will aim to identify these where possible as part of the conceptual model.

e. To the extent possible, multiple fisheries based on the same gear within the same region should be avoided, unless needed given difficulties in replicating tag returns. f. The Hessian matrix should be explored to assess not just the variances of the parameter estimates and the derived variables, but also which parameters may be highly correlated. Some further suggestions for model simplification were: g. Consider a model based on data for only the equatorial areas (regions 7, 8, 3 and 4) modelled as a single area and compare its results with an equivalently parameterized Stock Synthesis model. The model would be structured as "fleets-as-areas" with the fleets selected using, for example, the regression tree approach outlined above. h. Allow juvenile movement rates to differ from those for adults (which may be set to zero). i. Tag mixing issues: A fine-scale movement model would be useful for defining the time it takes for tagged animals to fully mix into the population within a region and experience the same probability of recapture as untagged fish in the region. A new approach was noted that models the spatial-temporal distribution of tags using advection diffusion models and the spatial distribution of the untagged population using spatio-temporal models is being developed for skipjack tuna in the eastern Pacific Ocean (Maunder et al., 2021; Mildenberger et al., 2022). However, methods to integrate the information from the analysis into the stock assessment model need to be developed.	
Representing uncertainty: The Panel noted that the construction of an uncertainty grid remains a state-of-the-art way to synthesize uncertainties that cannot be captured in a single run of an assessment model. They noted that the grid (and the a priori weights) assigned to the levels of each factor should be determined by the analysts who conducted the assessment. They provided commentary on approaches for weighting grids models, but no definitive recommendations. The panel recommended SPC staff attend various upcoming workshops on this issue and recommended deferring final decisions on summarizing uncertainty until after these workshops.	Characterisation of uncertainty in stock assessment outcomes, particularly as they pertain to provision of management advice, is an ongoing area of research. We agree that an uncertainty or ensemble model approach is as an important part of an appropriate characterisation of stock assessment uncertainty however, further consideration is required on approaches for inclusion/exclusion and weighting of models used to characterise uncertainty model grid ensembles. The SC18 has put forward Terms of Reference to review recent efforts to improve uncertainty characterisation (e.g. southwest Pacific Ocean Swordfish and blue shark assessments), with addition of a general review of methods applied more broadly across WCPFC stock assessments. This review would likely be delivered at SC19 so any recommendations would not be available for the 2023 assessments. We will also consider the discussions at the workshops suggested, and staff will attend the Tuna CAPAM in New Zealand, ideas form these forums will be discussed at the PAW – we hope these discussions might provide some useful and 'feasible' approaches to incorporate into the 2023 assessments of yellowfin and bigeye tuna.
Model diagnostics: The panel note that the push by SPC to develop more transparent dashboards to easily diagnose model results for different configurations is an excellent step in the right direction. The panel made note of several suggestions (pages 12-13) to enhance diagnostics.	In preparing for the 2023 stock assessments, we are continuing to build the content of the Shiny apps. for displaying diagnostics and key assessment results, including new diagnostics suggested by the peer review panel. The Shiny app. approach will become the main platform for displaying and comparing

	model outputs and diagnostics and will be available for external scientists to view.
Recent MULTIFAN-CL model developments: ***Catch conditioned approach:*** *The Panel endorses this approach (referred to as catch-conditioned model), which should simplify the models and ideally help to achieve a positive definite Hessian matrix.* ***CPUE likelihood:*** The Panel recommends a second option be developed where the variance of logCPUE is the sum of the square of a pre-specified CV and an overdispersion variance. ***The orthogonal-polynomial parameterization of recruitment:*** The Panel was concerned that this approach added yet another dimension to the model specification process and recommends that it only be used for data-poor situations or for the earlier years for assessments of data-rich stocks for which there is often limited information to inform estimates of recruitment. ***The Dirichlet-Multinomial distribution for length- and weight-composition data:*** The Panel endorses this approach but recommends that it be considered alongside the robust normal distribution and McAllister-Ianelli tuning. The Panel notes that all methods for weighting composition data depend on 'stage-1' sample sizes and emphasizes the importance of specifying these correctly.	The panel's endorsement of the catch conditioned approach supports the use of the method for the upcoming assessments. We acknowledge the panel's caution on the use of the orthogonal polynomial recruitment, which we don't expect to apply in the yellowfin or bigeye assessments. Improvements to the CPUE likelihood will be included as part of the MFCL development workplan for 2022/2023. We will test the application of the Dirichlet-MN for the 2023 assessments, considering the comments on the determination of the input sample sizes. There is considerable work involved in MFCL development if we were to implement all the recommendations from this review and we will need to be selective in what we attempt for the 2023 assessments. We are also in a phase of review and future planning for stock assessment software requirements, and this is also being factored into how we prioritise further development work on MFCL, including recommendations from this peer review.
Future research areas: the panel provides a list of future research areas to consider (pages 13-16). We do not address these all here, and many have been noted in the above sections of this table.	For the future research areas listed (pages 13-16) and other outcomes of the review, we consider the key focus areas for the coming assessments will include: • improving the model fits to size composition data. This may benefit from applying size-based selectivity rather than age-based selectivity, dependent on additional MFCL development work. The work on size composition data will also consider alternative initial sample sizes and data weighting approaches. • improving the reliability of the abundance indices, especially in how they are applied to inform the model on relative biomass among regions. • procedural practices such as documenting the bridging analysis, model assumptions and settings, and the enhancement of model diagnostics. Model convergence criteria will be also considered. • review of model spatial structure, including the life-history conceptual model and consideration of spatial patterns in size composition and or CPUE. • developing checks on biological plausibility including recruitment distribution and movement patterns with age.

	• consider alternative age-based movement hypothesis, and test sensitivity to different assumption on movement with age. • we will consider the options for treatment of tag mixing assumptions; however, we do not expect to be able to develop a similar simulation model as applied in the recent skipjack assessment in time for the 2023 assessments. Analysis of tagging data will rather focus on how this data can provide information on movement patterns that may support movement hypotheses. We will attempt to reduce the occurrence of tag reporting rates on bounds. • the approach to characterising uncertainty will be reviewed based on recommendations of the various workshops on this topic, notably the Tuna Assessment Best Practices CAPAM workshop in New Zealand in March, and the external review. • the future research recommendations noted by this peer review will be considered when updating the proposed Tuna Research Plan in 2023.