Title: Review of IOTC YFT & BET Assessment in 2019.

Author: Rishi Sharma, FAO, *Marine and Inland Fisheries Division,*
Food and Agriculture Organization of the United Nations, Viale delle Terme di Caracalla 00153 Rome, Italy.

Summary:

Different approaches were examined for assessing YFT & BET in 2019. A large effort was made to address issues identified in 2018 and the analysts should be commended on that. With respect to YFT, assessment examined in 2019, substantial issues relating to data quality were examined. Various assessment methodologies were examined and concluded that the stock continued to remain overfished; this includes a continuity analysis from 2018; however few models did not indicate overfishing trajectories were present, but more time needs to be spent examining these models, and weighting issues across models, and the most appropriate use of tagging information Some diagnostics indicate that information content in indices and length composition is limited and fail under numerous hypothesis examined (runs test and hindcasting tests). However, a much more extensive section on diagnostics was presented in 2019 as compared to 2018. Issues on high-grading were dealt with appropriately, but spatial and piracy issues need further examination for the standardization of CPUE. Overall, the process was transparent, and issues were briefly discussed relevant to uncertainty in the assessment results.

For BET assessment, the model appears to be correctly specified with no issues of finding a global minima. Models examined had some issues with hyper-depletion hypothesis that were discounted as the weight of evidence is that there is a decline in biomass in Area 1. Model diagnostics were performed extensively, and there appear to be no serious retrospective patterns. There should be simpler models examined for BET to corroborate the base/reference sets of assessment. Data inputs with respect to tag weights should be examined, and an understanding of CPUE declines in Area 1 need to be addressed. Issues of catches for the PS fishery need to be addressed, as these have large implications on stock status for both YFT and BET species. Issues of dome-shaped selectivity and plausible effects were examined. These have large implications on the assessment. As with YFT, the process was transparent, and issues were briefly discussed relevant to uncertainty in the assessment results.

A key limitation was that insufficient time was available to examine both data and assessment issues on multiple species at the meeting. If we could discuss model resolution and data before the meeting by species, additional time would be available to discuss further refinements in the assessments.

Keywords: Integrated assessment, CPUE, likelihood, data weighting, diagnostics, retrospectives, jitters.

Introduction

The WPTT was held in Saint Sebastian, Spain between 21st October and 26th October, 2019. The participation at the meeting included representatives from CPCs involved in the Tropical tuna fisheries (Taiwan,China, EU.Spain, EU.France,Sweden, Japan, Pakistan, India, Sri Lanka, Iran, Kenya, Somalia, Thailand, Indonesia, China, Australia, Sweden, Mauritius, Maldives and South Africa). Numerous other NGOs were present; CSOs (OPAGAG, ORTHONGEL, PEW, WWF-Pak were also present). This report addresses various issues that are important to WPTT and other issues being dealt with at the WPTT. Extensive work conducted on 2 complex assessments inter-sessionally, YFT and BET. All comments here are to help improve the process and assessments in the future.

1. ***Evaluate the adequacy, appropriateness, and application of data used in BET and YFT assessment.***

   Similar datasets (YFT has more artisanal fleets and datasets) are used for fitting purposes. Issues common to both BET and YFT are identified here, as the issues are similar and fleets involved for standardization purposes are the same. Issues on length composition and size selectivity (related to high grading) are discussed briefly, but more thought and analysis focusing on this and spatial coverage issues in Area 1 for assessments need to be addressed. The following areas are addressed in detail:

i) Four pieces of information are normally used in the assessment; they are the catch data, the length-composition data, the abundance indices, and the tagging data. Catch data had been examined carefully by each CPC (and the Secretariat), and all issues related to them are discussed by the Secretariat. Note, issues of species composition have been identified for the PS fleet. This is extremely important as it has implications on all 3 tropical species, namely Bigeye and Skipjack catches have increased in proportion as compared to the previous years, and Yellowfin has dropped in proportion as compared to previous years. These have implications on fishing mortality on all 3 species, and particular attention needs to be made to understand why these changes have occurred. Most of the studies from the EU fleet seem to contradict the reported catches, and the secretariat should examine this issue in detail. Primary issues relate to the large uncertainty in the data, and how this would be propagated in the assessment. Issues with catch reporting from longline fisheries in the 1990's were not discussed extensively, but coverage was known to be less than 10% in some years for log-book coverage in the Indian Ocean. In addition, issues with length-composition of the other and smaller fleets also need to be examined. Issues also of uncertainty in unreported catches is problematic. The secretariat makes estimates of catches and uncertainty on these catches, but how good these are is never debated. Alternative plausible catch series could be examined both within the context of the assessment and the MSE's.

ii) With regard to the abundance index data used in the assessment there were issues for each of the fleets and approaches identified:

- Purse Seine: Regarding how the purse seine CPUE standardization was done. The CPUE did not account for technological change which is a large factor that needs to be accounted for with this fishery. It was also not clear what the unit of effort used indicated (catch per free school set is cpue, but how many sets were made, what is the search time and how do they account for efficiency when they have high tech satellite technology to help fishing is never addressed), and there was also an issue with real versus set related zeros as a zero could just indicate a failed set even though fish were available under in an associated free school (note operational definitions such as large yellowfing associated FAD sets should me made clear). While a joint CPUE WG has examined some of these issues in detail, there are still numerous issues that

have not been dealt with. As such, the information content in the PS indices are limited, and have limited influence on the assessment, as indicated in WPTT-48 and by the assessment conducted in 2019 on both YFT and BET.

- Longline Combined Series: Note, the approach Hoyle et. al. (IOTC-2019-WPM10-16_0) used has a sub-setting algorithm which may influence the outcome, as well as the weights/regional scaling factors by area (IOTC-2018-WPM09-13). This was identified in 2018 and a similar approach was used in 2019, without examining how these may affect the outcome. Other issues not discussed were the coverage in recent years for some fleets have dramatically declined, and with a declining coverage, violation of certain assumptions on similar declines in other cells maybe unrepresentative. Walters (2003) discusses this in detail. This would have a large implication on the assessment (see Fig 1 below demonstrating how the CPUE may change based on different spatial interpolations used). While this has been known for a while, not much changes are made to how the CPUE is conducted.

  Issues on hyper depletion in Area 1for BET, and low abundances in YFT: These areas are poorly fished/covered by LL effort now. Historically, there was high effort and high catch rates for both species, which is now reduced primarily because one of the main fleets (JPN) no longer fish there. Given that some other approaches or hypothesis would be useful to examine using Walters (2003) or Campbell et. al. (2015) approaches as this has a large influence on the assessments. Taiwan, Province of China, has increased coverage and effort in these areas, and as such would be important to include in the standardization process in future years.

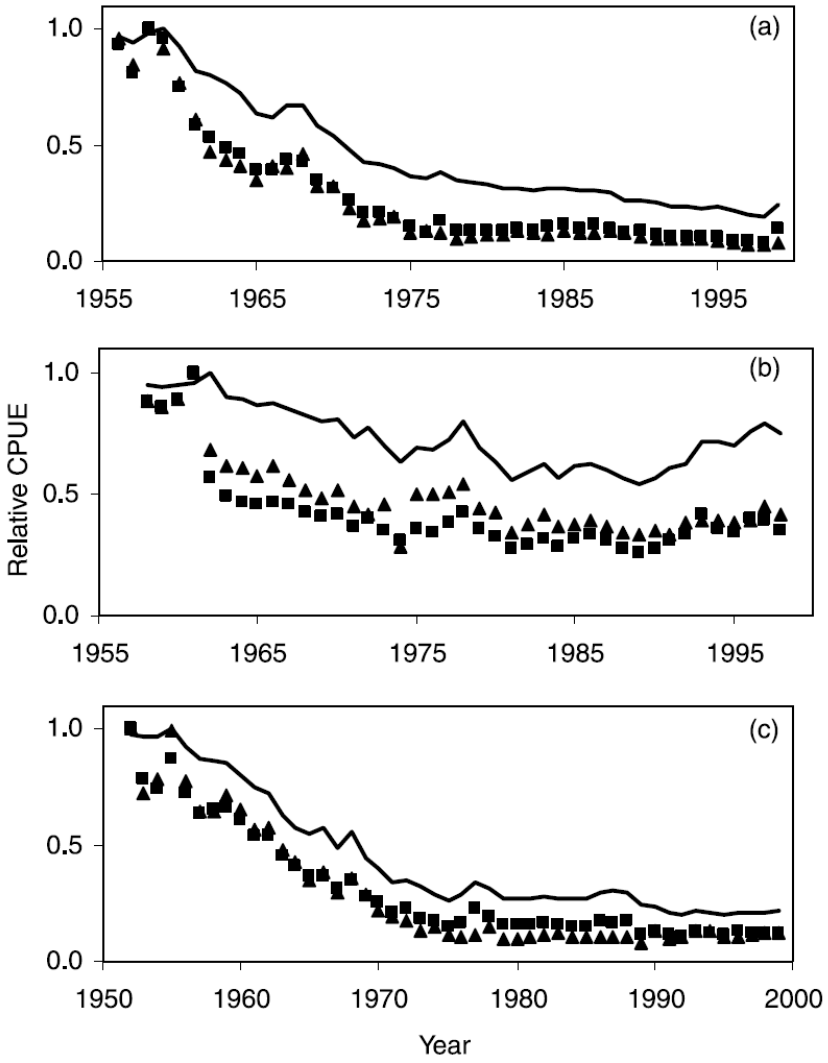

Figure 1 (from Walters (2003, CJFAS). Alternative hypotheses for how to fill cells with no coverage versus using the standard approach assumptions used for spatial extrapolation (Mean catch rates in over observed cells is always lower than the extrapolated values using coverage from last observations in that cell). Full spatial (solid line) assigns mean of first three observed catch rates to each cell for years before it was first fished and the last observed catch rate for years after it was last fished. Restricted spatial (block) is the mean catch rate over only those cells that were actually fished each year. Ratio (triangle) is simply total catch summed over all cells divided by total effort.

- **Other LL Fleets (Japan, Korea,Taiwan,China and China):** While this exercise is useful for characterizing fleet characteristics, it creates some confusion as to why we would care about a fleet specific CPUE versus a common CPUE series for the entire IO using all 3 fleets since a decision was taken a while back to use a common CPUE across all fleets. Issues identified above on coverage reduction, clustering and sub-setting are all influences on this analysis. Other fleet specific points are shown below:

  JAPAN CPUE STDIZATION

  1. Issues of common effects across fleets is examined.
  2. For a continuity analysis, it is important to account for and hence have the data from past approaches overlaid with this.

  KOREA CPUE STDIZATION

  1. Effort increasing over time and CPUE declining dramatically.

2. Look at common effects across fleets with JPN and Korea as stated above.

TAIWAN,CHINA (TWN,China) CPUE STDIZATION

1. Effort dramatically reduced in eastern IO. Hence how representative is the data in recent years can be an issue?
2. Issues of TWN,China fishing in coastal EEZs?

China

1. A new series was brought to the IOTC from China that has coverage in areas that has seen fleet reductions from both Japan and TWN,China.
2. Trends in BET appear to mirror what is seen in global standardization as well as in Japan and TWN,China.
3. It is important to add these data to the global CPUE standardization.

iii) The length frequency data were appropriately categorized and analysed for the fleets. However, not much time was spent on discussing why there were changes in one of the major fleets with length frequency data (possible issues of high grading after 2003) and implications on both the CPUE, and data used to infer recruitment. The minimum criteria as set by IOTC standards seem appropriate for representativeness for the fleet length-frequencies though how those related to the fleet stock compositions were not discussed. Issues relating to changes in length-frequency data for some of the LL and PS fleets seem to contradict the assessment results (possibly due to down weighting the length composition data). In addition time variant dynamics could be explored but largely ignored in this assessment. Shared selectivity modelled in the LL fleets may not be appropriate as there are subtle differences in how the fleets operate in different areas and can thus be modelled separately. Fits to those data could possibly improve if this is done.

iv) Tagging data from small scale tagging seemed problematic to use (recovery less than 1% overall), though examining the effect of this on assessment is not presented and should be assessed, before being discounted. The Maldivian and other small scale tagging, if it had been done correctly could give us information on movement and dispersal from other areas which would add the contrast and directional movement from other areas that is currently missing. At a later date, some additional analysis to examine what went wrong here, and why this data was discounted, would be important to understand. Again, the analysis of tagging data independent of the model needs to be examined. While there was a IOTTP conference in 2012, not much more has been done since then, and issues related to the following need to be addressed in detail: i) mixing periods and spatial scales, ii) growth and mortality, iii) tag shedding and mortality and iv) reporting rate estimation.

Overall adequacy of data used in assessment

Note that all assessments depend on the quality of data used. It is important to account for the uncertainty in the data, and examine sensitivity to alternative assumptions. The data used here is as good/bad as any other RFMO, as far as quality goes for use in the assessment. However, of particular concern is the catch information; as a majority of the catch are estimated and the model uses this as known (in the case of YFT the artisanal fleet quality of data is poor as indicted by the

IOTC secretariat as a lot of these estimates are calculated; in the case of BET issues on PSLS catches are probably estimated incorrectly, and need to be better reported; corrections were made on the fly based on estimated proportions from previous years; while this was a quick fix, there are obviously inter annual variations in these proportions and using the raw data is a better way to do this; **IT WOULD BE A HIGH PRIORITY FOR FLEETS INVOLVED IN THESE ISSUES TO PROVIDE AN ESTIMATE USING THE CORRECT DATA using a data prep meeting before hand**, as these have implications on other species encountered by this fleet). In addition, the LL CPUE has a large influence on the assessment and the data from 2007 onwards is probably not representative for the large drop in Area 1 (for YFT, and peak with subsequent drop for BET), the LL size frequencies are also problematic and create conflicts in the model fits (this latter issue is observed in both the Atlantic and Indian Oceans which may give some credibility to the fact that there maybe some high grading of smaller fish encountered by the TWN,China fleet). More time needs to be paid to details and examinations made to whether these are real or artefact of the data/problems in the standardization. While attempts were made to address these issues, the overall estimates were minor and maybe a revised estimate of high grading should be made for the all fleets (currently only TWN,China is corrected, but examination of this issue on other LL fleets is important.

In addition a meeting with CPC's to understand inconsistencies/changes in Length Frequency (LF) samples are important as these have large implications on the assessment. While the relevant CPCs meet for the standardization, I think further efforts need to be examined as to whether the data and spatial coverage by areas are representative in recent years especially due to effort and range restrictions of the LL fleet after piracy in the NW Indian Ocean (this was not examined in detail in 2019, as the focus of the work was on effects of high grading based on size sorting of the catch). In addition, it is extremely important to add Chinese LL fleet activity as they may cover the lack of coverage from Japanese and possibly Taiwan,China fleet that may improve the coverage in Area 1, as well as LC data for the LL fleet in Area 1.

2. ***Evaluate the adequacy, appropriateness, and application of methods used to assess the stock and if appropriate recommend alternative approaches to be accomplished in the future.***

We break this into 2 pieces, one focusing on YFT while the other works on BET:

YFT Issues

Two possible approaches were examined for the YFT assessment in 2019 and these should be sufficient to examine a range of possible options for the assessment; my comments will be addressed to each of them separately. In addition, I have summarized some basic information that maybe useful in examination for introductory purposes:

i) Examining simplified methods to assess signals in data- **NOT DONE**

Using simplified catch-curve analysis, using LC data by fleet, it would be easy to assess whether there are signals in the data suggesting that selectivity is dome shaped or mortality is U-shaped (based on ages of catches by fleet over time). Such examples are useful to assess if there is any signal in the data, and appropriate assumptions to be used in assessments. These approaches could be used to provide hypotheses for selection pattern for use in SS and trends in F, as a starting point. While these are standard exploratory data analysis techniques, none were really explore or presented in 2019-WPTT-21.

ii) Surplus Production based assessments

2 assessments were examined using this approach, namely JABBA (Winker et. al. 2018) and MPB (Kell et. al. 2015). One uses both observation and process error; the other uses only observation error. Both models examined indicated that the dynamics show a declining trend, and that the stock remains overfished with overfishing still occurring. Diagnostic tools presented can

be compared across different modelling platforms that were presented in the paper using JABBA. Some of these diagnostics were examined by alternative assessment models, but this analysis was not presented systematically across all models (4 area models were not examined like the 2 area models).

iii) Integrated Assessments (SS3).

Background Material/Model Specifications: Fishery resolution/specification indicates that we have 25 fisheries primarily PS (log and free school though the latter is becoming smaller over time), the LL fleets by different areas. While the fishery structures have not changed from the previous assessments, it may make sense to split some LL fleets into flagged vessels as currently they are assumed to have the same selectivity which may not be the case. This is something that has simplified the area structure, but the fishery resolution still remains the same.

The PS fleets could also be split particularly in areas where there is a bimodal distribution of catch into 2 bins, small and large. In addition the area stratification may need to be split out in Northwest as it was before around Oman, and the current 4 areas (total of 5 areas). The split from 5 to 4 areas with no CPUE series maybe causing problems as the current Area 1 is really driving the assessment and a drop in abundance there has a large influence on the overall assessment. Hence, more time needs to be devoted to the reasons in the drop and if this is a real artifact or is an issue with the procedure and lack of spatial coverage by the fleets. In addition, splitting the catch into 2 areas with a differential treatment as shown in Langley et. al. (2012) maybe more appropriate as movement could be assessed there (Figure 2&3 from Langley et. al. 2012).

The 5 area model analysis was not conducted, as this really does provide the value of using the tagging data. Instead the 2 area model without any tagging was used, and justified by diagnostics and that it mimics similar dynamics of the previous assessment. The logic on the latter is probably not great, as if model misspecification gave us a trajectory, we are mimicking the same mis-specification with a new structure. I would recommend using the original structure in Langley et. al. (2012), as well as no area model as another extreme.

Parameters in the 4 area model where we had hit parameter bound issues should be examined, and by resolving those possibly the model mis-specification and issues with instability in jitters could be resolved.

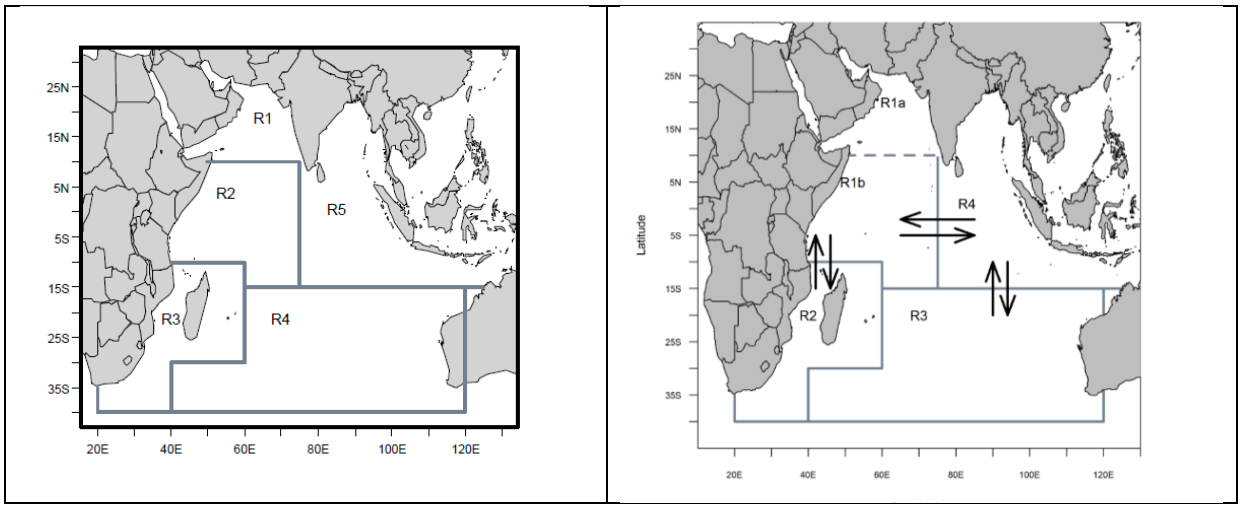

Figure 2: Spatial coverage from 2012 vs 2018

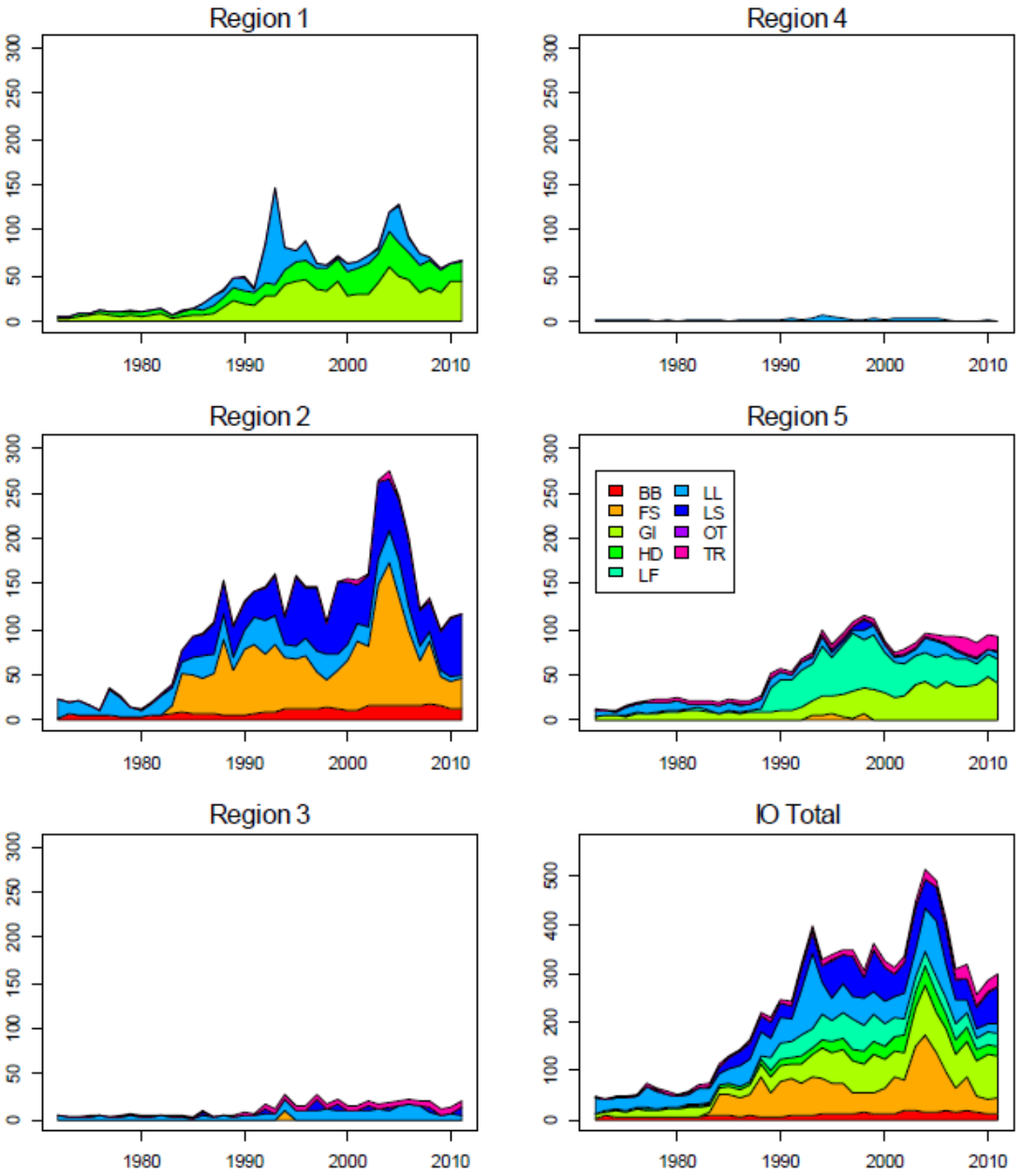

Figure 3: Catch Distributions by 5 areas (from Langley et. al. 2012).

Iterative reweighting approaches were examined in the 2 area models and different weighting mechanisms gave different results. The Ianelli data weighting method was used based on general stability compared to Francis (2011) reweighting approaches. However, Francis weights are normally the better approach when we have conflicts in data, possibly further down-weighting the ESS after iterative reweighting is proposed. This is a critical piece that is recommended and follows approaches suggested by Francis (2011). Alternatively, estimating selectivity using length frequency data, and then fixing it without using it in the assessment, i.e. completely down-weighting the length composition data. Data after a certain period on LL was not used (after 2004/ 2014)

The main issues identified in the 4 area runs presented were:

1. Examine in more detail the effectiveness of tags on the assessments in conjunction with other datasets in the assessments. While addressing some conflicts in the new 4 area assessment, this was completely discounted in the 2 area assessment. The 2 area assessment, is assuming no movement after recruitment, and some of these assumptions seem problematic.

2. Profile likelihood plots on R0 were done in the 2 area, and 4 area models. These are useful in identifying conflicts in data, but while some of these models were corrected for 2 areas, the 4 area model was not corrected.
3. Jitters examined indicate model convergence issues for the 4 area, but seem good for the 2 areas.
4. Multiple diagnostics indicate very similar model performance that shows both 2 and 4 area models are likely. The 4 area model uses tags so is more useful in describing movement and spatial structure useful for yellowfin.
5. Length frequency samples collected could be an artefact of sampling in some fleets in the latter years. As such, it may make sense to estimate selectivity and then not use the LF data as it wouldn't influence the estimate. This is still relevant and could be used, unless we think there is information in the length composition data to estimate recruitment..
6. Natural Mortality/Growth/Selectivity interactions need to be examined more carefully in these assessments as they have a large effect on the outcome of the assessment. In addition, M could have been estimated in the model with tagging data, but was ignored here. In the past (Langley et. al. (2012) did estimate the age specific M parameters over time, and examined their effect in the assessment but was missing here. Finally, using a grid structure provides context but examining these in more detail may indicate what the more plausible models exist within the established grid. Grid structures were examined along with differential spatial assumptions, and more coverage on structural uncertainty is examined.
7. Multivariate normal approximations developed by Henning et. al. (2019) is useful to give within model and across model uncertainty. The approach captures the true uncertainty and can be used to provide stock status advice for different catch levels as well.

It is **RECOMMENDED** to down-weight LF information and fit more to the abundance indices. The R0 was affected by length-composition and tagging data in contrast to indices in opposite directions. Further, additional analyses were made to fix selectivity based on fits to the LF data, but then to only fit the model to the CPUE series.

Bias correction issues on recruitment (Methot and Taylor 2011) need to be addressed. It is not clear if it is currently being done, and if so correctly. More time to examine this issue in detail should occur in the future.

OTHER CRITICAL ISSUES THAT WERE PARTIALLY ADDRESSED:

a) **Alternative CPUEs :** Examining catchability changes for the LL fleet in Area 1 could be examined, as something fundamentally different occurred in the LL fleet after 2007-2011 (piracy period). This is evident from the hindcasting exercise presented by Dr. Kitikado at the meeting. Examining different regional scaling factors for the CPUE could also be another approach to use as these have a large influence in the assessment. In addition, more effort should be made to find some fleets like the gillnet fleet (Pakistan), Maldives fleet (PL) and others (additiona LL datasets existed for some fleets like India, 2015 WPTT) where we can verify/calibrate the trends shown by the joint CPUE standardization. These would give more credibility to some of these dramatic drops seen in the standardization that effects these analysis and have large implications on the assessments. Alternative catchability was examined but dropped in this assessment as other structural uncertainties were examined (spatial structure).
b) **Data weighting and conflicting sources of information for assessments:** Based on the conflicts in length frequency, tag data, and index of abundance in the datasets, using

some down-weighting of the LF data or ignoring it entirely is recommended. This was done in 2019, and also additional downweighting and exclusions of data was done in 2019. As Francis states in his paper (2011) to use 3 principles in fitting models to data: "Principle 1: Do not let other data stop the model from fitting abundance data well; Principle 2: When weighting length composition data, allow for correlations; and Principle 3: Do not down-weight abundance data because they may be unrepresentative." Since tags appear not to be randomly mixed in the model, downweighting their effect on the likelihood is also recommended which has been done. However, given the weight they have on biomass (B_0)scaling (of ??) understanding how they interact with other components of the likelihood, and detailed examination on the data external to the model would be important in future iterations. The complex spatial structure was used to accommodate the tags (Langley 2012), so discontinuing that approach and having two areas is not preferred. In that case a one area model is more defensible especially if we are not modelling movement or exchange across these populations. These are recommended guidelines to be used. Final runs should examine fits to the CPUE with effort creep separately than the PL fishery, as these may give an entirely different picture of the stock. Other diagnostics such as hind-casting could shed light on which pieces of data have information and which do not, thereby examining different weighting schemes based on diagnostics.

c) **Dealing with Uncertainty:** In addition, when using forecasts, MCMC based projections could be examined for the reference run. These runs could be compared to the multivariate normal approach to make sure that the uncertainty bounds across these platforms are comparable.

BET Assessment

Examining simplified methods to assess signals in data- NOT DONE

Using simplified catch-curve analysis by fleet, it would be easy to assess whether there are signals in the data suggesting that selectivity is dome shaped or mortality is U-shaped (based on ages of catches by fleet over time). Such examples are useful to assess if there is any signal in the data, and appropriate assumptions to be used in assessments. These approaches could be used to provide hypotheses for selection pattern for use in SS and trends in F, as a starting point. While these are standard exploratory data analysis techniques, none were really explore or presented in 2019-WPTT-21. Simple analysis like this would inform the assessment structure to be used in the current year

ii) Surplus Production Assessment (JABBA Assessment):

A surplus production assessment was conducted and diagnostics indicated that the stock is not overfished nor experiencing overfishing. Simplified assessments are a good check against a simplified assessment, and 2019-WPTT21-32 indicates that the stock is not overfished or experiencing overfishing. The bulk of the stock status density remains in the 1st green quadrant. These profiles are a good diagnostic check to demonstrate whether the complex models remain within the uncertainty envelope of the simple models. If they were differences, it is likely that model mis-specification is occurring. Similarly MPB Model analyzed provided similar conclusions on the stock, However, due to limitations on size selectivity issues in these Biomass Dynamic model, JABBA-SELECT would be the way to proceed in the future.

iii) SS-III Assessment:

A comprehensive assessment was run in 2019, 2019-WPTT21-61; this model presented a set of diagnostics, and seems to have reasonable performance. However, the following issues were important to consider:

1. Spatial structure and tag mixing issues make this model drive to a lower Biomass, i.e. scaling issues to drive the biomass low. These are probably occurring due to incomplete mixing. 3 area models examined in the past (Langley et. al. 2013) discounted this data, and used one area with fisheries by area as the preferred approach. This was changed in 2016 as it was prioritized to use the tagging data. However, a thorough analysis of the tagging data needs to occur and assess whether it should be used internally in the assessment or to estimate life history parameters and fishing mortality, F.
2. Catch changes from PS fisheries in recent years seem implausible, given the fleet is still operating in similar areas and manner as before. Implications on catch changes and allocation have a large impact on the assessment. CPCs should examine this data, and decide what is the most plausible catch scenario, and correct this. This is a concern as the PS is one of the fisheries that is sampled extensively so if they are mistakes in reporting these estimates, then it could be a much larger problem for other fisheries and CPC's/catches at large.
3. CPUE Standardizations in Area 1N (Northern sub division) drive the assessments: The spike in 2011, and subsequent decline could be an artifact of the piracy/fleet fishing change. The decline is also creating issues (may want to use same approach as 2016 assessment, common signal for area 1 vs 1N and 1S). This may get rid of some of the patterns we see in these areas. Alternative hypothesis on catchability change could be equally plausible, and initial model runs accounted for this, but were discounted in later years.
4. Length composition were down weighted extensively; this helps inform selectivity, but not inform recruitment; and was done properly this go around.
5. Rather than use a common selectivity index, a Logistic Selectivity should be used in Area 1. Region 1 is shown to have the large fish. It was recommended to anchor the logistic selectivity in region 1 and estimate dome shaped selectivity for other 2 areas. The steep incline/edge of the dome shaped selectivity is highly unlikely and needs to be corrected in the future to use a smoother shape.

Due to insufficient time, not enough time was spent on diagnostics though the jitters and retrospectives on initial models presented indicated that the model was stable. Alternative structures without the tagging data could be examined in future years. This should be a HIGH PRIORITY for examination in future years assessments.

3. ***Evaluate the methods used to estimate population benchmarks and stock status (e.g., MSY, FMSY, BMSY, or their proxies).***

Reference points estimated are a function of the information used in the assessments (i.e. length frequency data, the abundance data (CPUE) and the catch data, as well as the tagging data). For integrated assessments, the selectivity estimated and the values used are critical in estimating the key reference points (MSY, BMSY, FMSY and relative levels of fishing wrt to these reference points). Most models examined had similar values for selectivity (whereas M and steepness were fixed), and as such using some assumed selectivity (estimated from the data) and fixed M and h, will provide consistent reference points across a number of model runs. However, some of the selectivity in recent years was predicated on the length-composition data, and these may not be known well; in addition examining alternative dome shaped selectivity for the LL fleet and its effect on reference points . These latter points on having differential selectivity by areas or different assumed shapes were examined for BET but not for YFT. Some of these issues should be examined in the future.

In addition, growth should be examined as to how it affects advice on MSY and current catch levels with respect to MSY. As such, some of the absolute measures may be inaccurate, but the relative reference points (B_{curr}/B_{msy} & F_{curr}/F_{MSY}) should still be a good indicator of

stock status. The estimates as such from SS are probably more reliable than surplus production models as they deal with selectivity across fleets and surplus production models cannot explicitly do so. However, given the problems with the data, the surplus production based approaches work just as well. Note, the use of virgin biomass (K) as a reference point shown at the meeting is useful as a reference point (some fraction of B_0 as a target and limit), as it remains independent of selectivity and its effect on MSY estimates, and is suggested here, and I strongly support this. WCPFC use this since it is both independent of steepness and selection pattern. SPR0 is multiplied by the recruitment each year to give a changing biomass reference point, and maybe a better alternative to use.

Both YFT and BET Assessments maybe underestimating MSY as models seem to be underestimating the sustainable yield as larger catches are not collapsing the stock. Of particular concern is that the MSY target for BET has changed dramatically (reduced by ~70K t from 2013) as the tags and area based assessments have downscaled the initial biomass thereby reducing the MSY estimates by a large factor. Additional work needs to be conducted to assess whether tags are properly mixed as these have large implications on the assessment.

4. ***Evaluate the adequacy, appropriateness, and application of the methods used to evaluate future population status, given the commissions objectives.***

YFT: No considerations were given to future population status with catch projections. I find this disconcerting as that's really what is most important. It's not where we were but where we are going. Given that we don't have another assessment for 2 more years some considerations should be given to this, unfortunately no discussion or time was spent on this. Note, a reference grid was built and projection advice could be provided based on corrections made to the last grid as well and presented at 2019-WPTT-21 by Dr. Winker, but was not adopted. There is a fine line between some and no advice, and as such it would be good to report on projections based on the last assessment, as issues with projections that occurred in 2018 were corrected.

BET: Deterministic projections based on the 18 models were conducted, and the projections indicate that the stock is probably being overfished but overfishing is not occurring currently. Stock trajectories do not make sense especially as catch biomass fished in the early 2000's were substantially larger in the past. The multi-variate lognormal approach makes sense as it makes projections easy to conduct and provides estimates of uncertainty extremely quickly without running extensive grids and projections. The current catch levels appear to be unsustainable, and need to be dropped by 20% to keep the stock above BMSY with greater than 60% probability in 10 years.

5. ***Evaluate the adequacy, appropriateness, and application of methods used to characterize the uncertainty in estimated parameters. Comment on whether the implications of uncertainty in technical conclusions are clearly stated.***

Assessments in general in 2019 had a large discussion on uncertainty and diagnostics to evaluate how well the runs did using likelihood profile and other techniques Structural uncertainty of multiple models were not examined due to time constraints using runs test or hindcasting approaches; however data-weighting or alternative area examinations (5 versus 4 areas in the integrated assessment for example, 2 were used, but 1 and 5 area models could also be examined particularly given that original assessments had 5 areas) need further thought and development. Finer resolution fishery structures should also be developed to incorporate some of the fleet characteristics which may differ by flag (LL and PS fleets have very different operations by flag), and asymptotic versus dome shaped selectivity could be examined. In

addition, when using forecasts, using either structural uncertainty grids with deterministic catch or MCMC based projections should be examined. The MVN approach was presented and can be done quickly, however results were not accepted for YFT from 2018 assessment reference grid even though the issues identified in 2018 were addressed by Winker et. al. but the same approaches were adopted for BET, indicating in some inconsistency on what is applied. It was possibly due to time restraints that the YFT projections were not adopted, but issues related to model instability due to high F's (infinite) and low biomass were corrected by applying a max bound on F and a minimum bound for B

The final runs decided on were determined on a very arbitrary basis. Again, a better way to proceed would probably be to discuss these in detail before the assessment (or at another meeting) and then proceed with a whole grid and a partial grid based on the larger grid. While inputs at the meeting are useful, every analyst would want something different which makes it tough for the primary modeler to do everything. The process thus needs to be streamlined and be more efficient in how the WP operates for inputs to the primary assessment.

For YFT, a key issue that was not examined carefully was the coverage issue on CPUE in recent years as LL fleet activity has dramatically declined between 2007-2011 (piracy) and after that as well. In addition even though the Korean fleet effort has increased, its only 3 vessels operating, and hence issues of representativeness could be examined. Also, examining issues with the old area 1 versus new area 1 (combining area 1 and 2). Two hypotheses could represent this, i) standardization done for the period 1979-2017 is done correctly and the decline is real after 2007, ii) alternatively the catchability changed after 2007 due to different fleet*area structure/interactions, and this may not be representative of catchability and a catchability block could occur after 2007.

Similar issues on BET CPUE standardization in 2019 were evident after 2011-2012, and initial runs did examine alternative q hypothesis, but were later discounted (I think those are plausible runs and should have been part of the reference grid, but the group proposed alternative approaches). The spike in catch rates with subsequent decline after that appears to drive the overall assessment trend. This should be examined carefully when this is done in subsequent years. If fleets are modelled separately, possibly using fleet specific CPUEs could be applied as well.

6. ***How did the assessment inform the HCR and allowable TAC? Was the process well thought out?***

Not relevant as MSE in development currently for both BET and YFT. Development in WPM is ongoing and there are targets and control rules that need to be met soon.

7. ***Comment on whether the stock assessment results are clearly and accurately presented in the detailed report of the Stock Assessment.***

The presentations did cover most of these results adequately, but having written documentation available as well as an archived script for the model runs would help reviewers and participants follow proceedings at a later date (Amendment to the reports based on discussions at the meeting need to occur). Again, clear explicit requirements for assessments should be specified well in advance of the meetings, and deadlines set for all assessment documents to be made available before the meetings. While runs did occur at the meeting, for archiving purposes there should be an effort to document these runs at a later date.

8. ***Comment on potential improvements on the stock assessment process (CPC participation, transparency, objectivity, documentation, uncertainty characterization, etc.) as applied to the reviewed assessments.***

While extensive time was spent discussing alternative model runs and approaches, as well as the data at the meeting, I suggest the following steps to streamline the process; trying to do 2 BIG Tropical Assessments (BET and YFT is not recommended, maybe an update on YFT and a full assessment on BET, but doing both BET and YFT was not a suggested approach):

a) All datasets are made available to the modelers 2 months before the meeting.
b) Clear write-ups are made available on all approaches used in the assessments at least 2 weeks before the assessment meeting is held.
c) All approaches are discussed on the 1st day, with all additional runs (grids set up for the analysts on the second day)
d) All new results/approaches are presented on the 3rd day as 2nd day used for analysis (other business is covered in day 2 of the meeting). Recommendations on stock status and projections completed by 3rd/4th day after the final set of runs is agreed.

Alternatively, a week with a smaller group like (MSE small WG) work on data issues (like CPUE WG) and assessment issues simultaneously. This group would vet enough models and plausible hypothesis a month or so before the meeting and then present a thoroughly vetted process for the WPTT.

CPC participation was limited primarily to the developed nations (EU, Japan and Taiwan, and the Secretariat and the CSOs). **More time spent at the data meetings clearing the data issues of developing coastal countries that have important fisheries on the species that is the target of the assessment would substantially improve this process (e.g. Pakistan, Iran and Indian fisheries as well as the Sri Lankan gillnet fisheries and datasets).** Reports available were limited and while some runs were archived on the IOTC website, some additional readme documentation should go with this so people are aware of the approaches and possibly could run them if needed.

These are just ideas to make it more efficient. Given the timelines the modelers were given, the job and approach presented was more than adequate. However, given the value of the stock and importance of the species in the Indian Ocean, more time should be given to the analysis (a possible solution would be that the Commission changes the standards for the reporting of statistics so as the 2 meeting plan can be set and data from the previous year are available in time for the assessment). This would mean more time should be spent understanding and preparing the data so analysts could complete most of the runs before the meetings, and examine only a few hypotheses at the meetings.

9. ***Comment on the adequacy of the work plan for the assessment and whether it was adequately addressed by the WPTT***

The work plan used was adequate. More time needs to be paid to quality control on datasets provided by CPC's as these can have a large impact on the assessment and sufficient time examining these data is warranted in the future. A reexamination of the tagging data is warranted, as it has a large influence on scaling on both BET and YFT assessments. As it currently stands, CPC data are used with some proofing (though approaches used need to be clearly documented and understood by the CPC's involved as the Secretariat does this uniformly). There are obvious short-comings in the datasets being used (e.g. Length frequencies

should not be used blindly, nor should the CPUE as they have a large influence on the assessment), and the catch data expansion methods need review (also use and adequacy of the tagging data are important). Even though a joint process on CPUE standardization is done, some large drops in CPUE that do not coincide with large ecosystem changes or fisheries effects need to be examined as these maybe biased low. The trend after 2012 in BET and low rates in YFT in Area 1 are key issues of concern. As stated before, I think there should be a separate data preparation meeting and analysis for the stock being examined in the assessments, so the data can be analyzed adequately by the assessment scientists and reports describing the approaches are made available at least a month before the meeting where the assessment is discussed.

10. ***Consider the research recommendations provided by the working group and suggest any additional recommendations or prioritizations warranted. Clearly denote research and monitoring needs that could improve the reliability of future assessments. Recommend an appropriate interval for the next assessment considering control rules or management strategy in effect.***

Some of the key recommendations were on biology and growth of the species, which were not examined extensively. Further work needs to be conducted on cross-validation to assess which is the most informative series by using a hind-casting approach (Kell et. al. 2016 to assess model performance in a predictive sense.

A paper by Kell and Sharma (2019 WPTT21-48) provided examples of diagnostics that might be applied more generally in the IOTC stock assessment process for model validation, i.e. based on prediction skill and runs tests, The models, however, were not intended to provide management advice but to provide insight about uncertainty about IOTC YFT population dynamics. For example, estimates of Surplus Production (Walters, et al., 2008) can provide a check of whether predictions of changes in biomass can be made reliably based on catch and current biomass or whether there has been non-stationarity in production processes, i.e. are dynamics driven by climate and oceanic conditions (IOTC-2019-WPTT21-24). This is important for the development of MPs in the MSE process.

.

Further work is required to understand the data behavior (drops in CPUE in Area 1) and discrepancy in the LF data across similar fleets operating in similar areas. The main recommendations are the following:

1) To examine the PS CPUE series used, and improve it based on similar exercise undertaken in the Indian Ocean on LL fleets (see Hoyle et.al. 2015). In addition, a meeting with the DWFN LL CPC's to understand inconsistencies/changes in LF samples as these have large implications on the assessment. Some of this work has been done, but further examinations on fleet activity is required.
2) To examine the data coverage (spatial extent) of the LL fleets over time and whether we maybe overly extending the data and assumptions to the latter periods.
3) One should fit to each plausible hypothesis separately (catchability change over time for LL fleets versus not, use of PS fleet CPUE or not) as these are alternative states of nature and alternative hypothesis that you are testing against. As such, we need to evaluate this separately and not combining these indices simultaneously for PS and LL especially. This is true especially for surplus production model approaches and models using one area.
4) As far as integrated analysis are concerned further examination should be conducted on the following items:
    i. Weight the model fits to CPUE series rather than LF observed in the fleets.
    ii. To examine Natural Mortality/Growth/Selectivity interactions more extensively as these are critical to the assessment.

iii. To examine the data weighting issues on tags, it is recommended that a through analysis on tag data outside and withing the model is done. This has a large influence on biomass scaling, and is very sensitive to mixing. Using the tagging data to design the spatial structure is also important.

iv. To make sure that uncertainty is accounted for accurately. Grid based versus MCMC based. One run versus many runs and grids (more thorough interactions should be examined so a larger uncertainty that accounts for biological effects and data effects and interactions). However, for a later period a more thorough examination using MCMC and a more expansive grid should be examined. These approaches should be tested with MVLN approach developed by Winker et. al. 2019.

4) Issues of local minima are a concern in these over-parameterized models. Using multiple diagnostics like RO profiles (information content in the data), jitter analysis (check for convergence and local minima issues), and retrospective patterns (ability of model to capture trends overtime). Although some checks were done at the meeting, insufficient time was spent on diagnostics that need to be accounted for at a later period. In general, for both YFT and BET, extensive effort was made to examine diagnostics.
5) Issues of spatial complexity; going back to the original structure of Langley et. al. (2012) maybe more appropriate as effort has moved back to the old Area 1 and the movement data was more informative using that as well as fleet structures made more sense in terms of separation of effort of fleets by area.

Given the stock status indicators from the alternative assessments, the stock is probably overfished and is likely experiencing overfishing for YFT. For BET, the stock is not overfished but probably experiencing overfishing. However, alternative hypothesis of catchability drop for LL fleets would give a very different outlook on the stock for both assessments, and down weighting or excluding tagging data would change the assessment outlook dramatically. A more thorough examination should be made on these changes and how they affect the assessment. Finally, equal weighting of all models is probably not a recommended approach. There should be a reference case assessment and then a plausibility bound with the sensitivity runs.

11. ***Other papers of relevance on WPTT***

Numerous other papers were presented, but the ones on CPUE standardization of all LL fleets on the Indian Ocean were important as they discussed issues that are of utmost important in how the series should be developed for future assessments on Yellowfin and Bigeye, and how we need to pay particular attention to certain discontinuities in the data, the issues of spatial resolution and weights to use in assessments, and the issues of length frequency data getting worse over time for some fleets. Other CPC papers on issues relevant to their jurisdictions were discussed, and have relevance to issues such as catch compositions and length frequencies for both species.

Overall Conclusions

The use of multiple approaches is important when assessing stock status. While different approaches were examined (ASPM vs SS vs JABBA), time was initially spent on diagnostics (jitters, profile likelihood, and retrospective analysis), but insufficient time precluded diagnostic evaluation for the final models used in BET, but some of the work was conducted for YFT 2 and 4 area assessments without any conclusive evidence of preferring one model over another, nor was there enough time spent on understanding why indices were behaving the way they were for the LL fleets, and possibly examining other hypothesis. Arbitrary decisions (giving equal weight to all models seems overly generous as some models should be more plausible than others) on what the final models to use for advice were developed without a thorough

analysis, as decisions made on the fly can have large implications on the assessment (tag weighting, growth and M assumed for YFT, and tag weights for BET). Length frequency data are particularly important for SS, and as such examining if these data are accurate is critical in the assessment, as they are currently being down weighted in both assessments.

Tagging data sensitivities also need to be examined more thoroughly, especially with regard to mixing (number of quarters to exclude), tag mortality and shedding rates, and over-dispersion parameters used. Currently, it has been pointed out that there are some critical uncertainties in both the CPUE data used in the assessment and the length-frequency datasets, and as such warrants further examination. Assumptions on tag release mortality and its effects also need to be examined in detail. These will all have a large effect on the assessment. In addition, for integrated assessments, it is critical to examine the data weighting issues and what drives the assessment. Francis (2011) points out that 3 principles are important when conducting an assessment, and these are: "Principle 1: Do not let other data stop the model from fitting abundance data well; Principle 2: When weighting composition data, allow for correlations; and Principle 3: Do not down-weight abundance data because they may be unrepresentative." This was attempted to some extent; however additional analysis is warranted on this issue. In addition, CPUE in recent years for some areas (Area 1) have large implications on both assessment and methods to deal with this with rationale could be examined in the future.

Overall, the process was transparent, and numerous issues were discussed. A key limitation was that datasets need to be examined and finalized with more lead time, so actual papers and analysis are available and discussed in advance of the meeting (possibly with a smaller group discussing data issues and fisheries resolutions that should be examined with enough lead time for the assessment analyst). If this were done, efficient use of time would be spent on discussing further refinements in the assessments rather than spending time making ad hoc decisions at the meeting. Finally, approaches dealing with uncertainty and projections were not given due importance, as normally after dealing with the assessment issues insuffiecnet time remains for dealing with projection issues, but as these are critical for stock status advice, and management advice that would sustain the long-term sustainability of the stock, additional time should be spent on these issues in the future (possibly intersessional papers should be circulated before the meetings so these items are discussed extensively at the meetings).

I would like to commend Mr. Fu, Dr. Agurtzane Ijurco, and Dr. Cardinale for doing an excellent job on both assessments along with diagnostics. Comments in this report are intended to be constructive and help improve the assessment in future years. Given time constraints analysts did an excellent job.

Acknowledgements

I would like to thank the secretariat (namely Dr. Chris OBrien, Dr. Paul DeBruyn, Mr. Fu and Dr. Fabio Fiorellato,) for giving me the opportunity to come as invited expert for WPTT 2019. I would also like to thank the Chair, Dr. Gorka Merino , vice chair Dr. Shiham Adam, and SC interim-chair Dr. Shiham Adam for conducting the Working group meeting efficiently.

References

Francis, R. I. C. C. 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. 68:1,124-1,138.

Gaertner, D. and Hallier, J.P. 2014. Tag shedding by tropical tunas in the Indian Ocean and other factors effecting the shedding rate. Fish. Res.163:98-105.

Geehan, J. Hoyle, S. and Herrera, M. 2013. Review of length-composition data of Taiwanese Longline Fisheries. IOTC–WPTT15.

Winker, H., Carvalho, F. and Kapur, M. 2018. JABBA: Just another Biomass Assessment. Fish. Res. 204: 275-288.

Hoyle, S.D., Okamoto, H., Yeh, Y. Kim, Z., Lee, S. and Sharma, R. IOTC–CPUEWS–02 2015: Report of the Second IOTC CPUE Workshop on Longline Fisheries, April 30th– May 2nd, 2015. *IOTC–2015– CPUEWS02–R[E]: 124pp.*

Hoyle, S., Leroy, B., Nicol, S., and Hampton, J. 2015. Covariates of release mortality and tag-loss in large scale tuna-tagging experiments. Fish Res. 163: 106-118.

IOTC (2005). Report of the Ninth Session of the Indian Ocean Tuna Commission. 15th session: IOTC Doc IOTC-2011-S15-R [E]. Victoria, Seychelles, Indian Ocean Tuna Commission.

Kell, L.T., Kimoto, A. and Kitakado, T., 2016. Evaluation of the prediction skill of stock assessment using hindcasting. Fish. Res., 183: 119-127.

Langley, A. Million, J., and Herrera, M. 2012. Stock Assessment of Yellowfin Tuna in the Indian Ocean using Multifan-CL. 2012 WPTT 14-38

Methot, R.D. and 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.

Walters, C. 2003. Folly and fantasy in the analysis of spatial catch rate data. *Canadian Journal of Fisheries and Aquatic Sciences*, *60*, pp.1433-1436.

Walters, C.J., Hilborn, R. and Christensen, V., 2008. Surplus production dynamics in declining and recovering fish populations. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*(11), pp.2536-2551.