Title: Review of IOTC YFT in 2018.

Author: Rishi Sharma, NOAA, NWFSC, Conservation Biology Division, 1201 NE Lloyd Boulevard, Suite 1100, Portland, OR 97232, USA

Summary:

Different approaches were examined in the YFT Assessment examined in 2018, however, conflicting data inputs, data weighting and catchability changes may create problems in the assessment. Issues of convergence/local minima need to be checked thoroughly with jitters. Other diagnostics such as profile likelihood techniques and retrospectives are important to examine. Overall in general this topic needs more time and coverage in the future, as currently the diagnostics examined were limited. In addition, length frequency data are particularly important for the age structured assessments used, and accuracy in these data is crucial to the inference, as large uncertainties still exist in this series (particular attention to ESS and how they influence inference are important). Data weighting issues were not examined extensively, and further work is warranted regarding this subject, as weights between length composition data, CPUE and tagging data can provide very different inferences on the population. Overall, 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 at the meeting. If we could discuss model resolution and data before the meeting, additional time would be available to discuss further refinements in the assessments. Preliminary analysis using hindcasting techniques suggest that the model has poor predictive power which is a concern, and may also be problematic as it has local minima issues. Finally, approaches dealing with uncertainty and projections were not given due importance, 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.

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

Introduction

The WPTT was held in Mahe, Seychelles between 29th October and 3rd November, 2018. The participation at the meeting included representatives from CPCs involved in the Tropical tuna fisheries (Taiwan,China, EU.Spain, EU.France, Japan, Pakistan, Sri Lanka, Iran, Kenya, Thailand, Indonesia, China, Australia, Sweden, Mauritius and South Africa), though some important countries like India were missing. This report addresses various issues that are important to WPTT and other issues being dealt with at the WPTT.

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

i) Three (possibly 4) 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. 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, but how good these are is never debated. Alternative 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, 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 a FAD.. While the indices presented seem appropriate, the methods used could be fine-tuned or improved and account for other measures of effort (PS) like search time and distance. There were some key issues with regard to confidence intervals of these series (negative values generated), and also the issues related to whether combining the data from free schools and FAD schools were appropriate as they maybe fishing on different size classes (evident in the bimodal structure of the data). This brings another topic that possibly this fishery should be split into two components in the model with different selectivity for each fleet.
- Longline Combined Series: Note, the approach Hoyle et. al. (IOTC-2018-WPM09-12_Rev1) 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). 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. Issues on catchability declines due to i) environmental influence (Marsac Paper IOTC-2018-WPTT20-XX) could explain changes in catch rates as decline of DO content on surface means that these fish go below 30m and thus may not be seen in the catch, ii) effort coverage that maybe unrepresentative, and iii) different cluster and sub-setting algorithms could provide alternative hypothesis and trends to examine, iv) regional scaling factors have a large influence and more care should be given to how these are weighted. Another issue of mix and match of approaches based on period covered. In addition, having a truncated series in 1979 could be mixing two different series and approaches; implying it would be better to have the longer series with clusters rather than the hooks as indicator for targeting. Finally, the drop in 2007 in Area 1 to a new equilibrium implies 2 possible hypothesis; i) the abundance has truly dropped and the stock productivity has declined, or ii) the standardization hasn't taken into account catchability changes that implies a new catchability and hence new fishery/environmental behavior or drivers on how the fishery interacts with the stock.
- Maldives Pole and Line Index: The Maldives PL fishery and index has a few problems identified:
  1. The trends in SKJ were not realistic. YFT may be more plausible. However uncertainty is large for YFT, Need to revisit the priors to be used in the standardization procedures.
  2. Compare this to biomass trends as well. Also, compare to what was previously done in 2013. Factor of 4 catchability change due to expert opinion maybe a little problematic. Either revisiting the expert opinion or making a change in how this is used.
  3. If it's a localized depletion in the Maldives, how do we reconcile this is the assessment (more a 3 box model).

- **Other LL Fleets (Japan, Korea and Taiwan,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 that common effects across fleets is examined.
2. For a continuity analysis 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 CPUE STDIZATION

1. Effort dramatically reduced in eastern IO. Hence how representative is the data in recent years can be an issue?
2. R1 no analysis shown. East region has lower catches and not include data from small or large vessels. Data only used from large vessels.
3. 1A combine with western Tropical Region. Hence, no index and maybe something could be generated there based on Taiwanese fleet data.
4. Cluster analysis combines species combination and type of variables to get an integrated set for targeting. Due to randomness in catching something else could be a problem, and hence makes sense using some other algorithms to see if these have effects on the overall signal.
5. Core areas for region specific effects. Choosing to see trends in all regions are same. Smaller areas and then wrap to other areas.
6. Difference between 2 trends. Joint indices are developed with HBF and TWN used clustering. Probably sufficient explanation for difference.

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. 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.

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.

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 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, and 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 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. 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 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.

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.***

Two possible approaches were examined for the YFT assessment in 2018 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 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 2018-WPTT-20.

ii) Surplus Production based assessments
While some papers were presented for simplified assessments, not enough thought was paid to these approaches. JABBA methods (Henning et. al. 2018) have high relevance especially for use in the context of an MP and HCR. The invited expert did examine this with Dr. Henning (attached figures), and the conclusions are slightly different than the main assessment (Figures 1-5). Some diagnostics like the following could determine which model is more appropriate, and preliminary research conducted by Kitakado, Kell and Sharma (in progress) indicates model parsimony often provided better predictive power and follows trends closer than over-parameterized models. The following approaches could be used

a) We should use a hindcasting/retrospective analysis for deciding on how good the models are. In addition jack-knife and the ability to predict missing data is a good diagnostic to run for determining relevant use of these models (see WPM Report from 2017, and joint TRFMO meeting presentation by Sharma, Kell and Kitakado).

b) Uncertainty estimation using different approaches (Bootstrap, MCMC, Hessian) could possibly be evaluated easily with these simpler models, and should possibly be used to examine the more complex models as well. The choice of how you estimate uncertainty has a big effect on the assessment especially for projections.

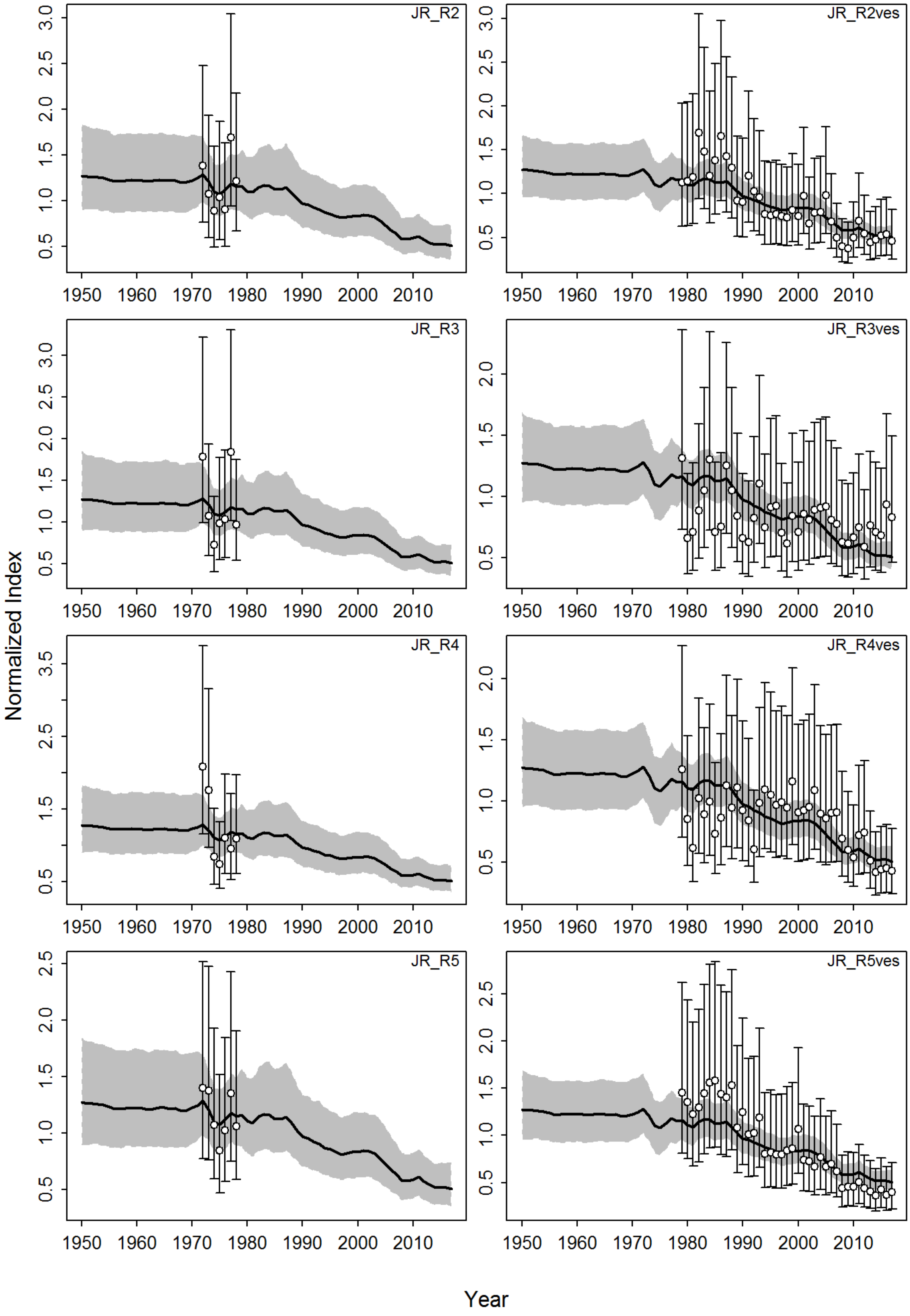

Figure 1: YFT stock trajectory with respect to CPUE series used by 4 areas.

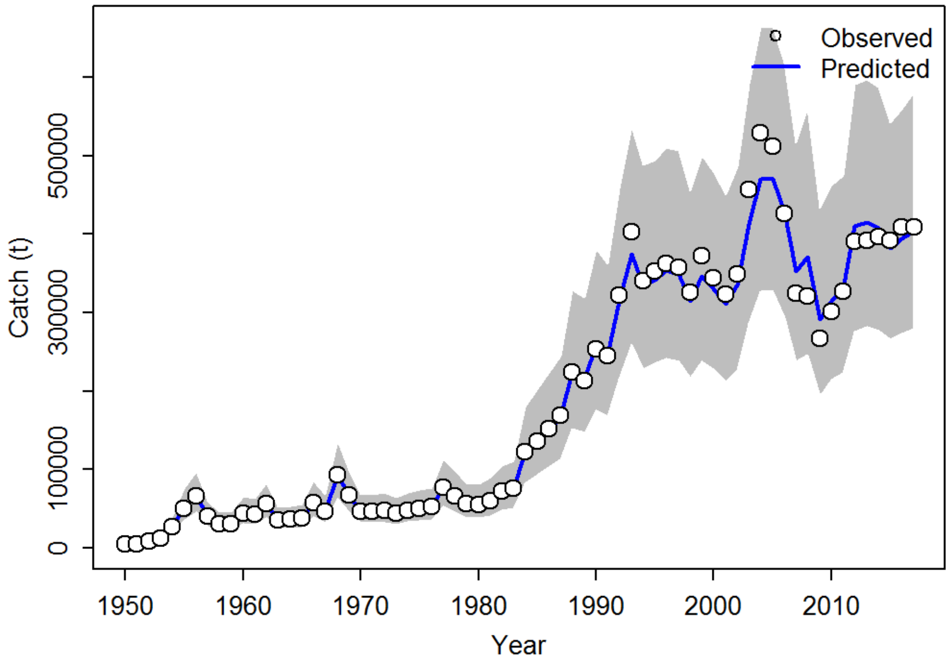

Figure 2: Fits to the catch series

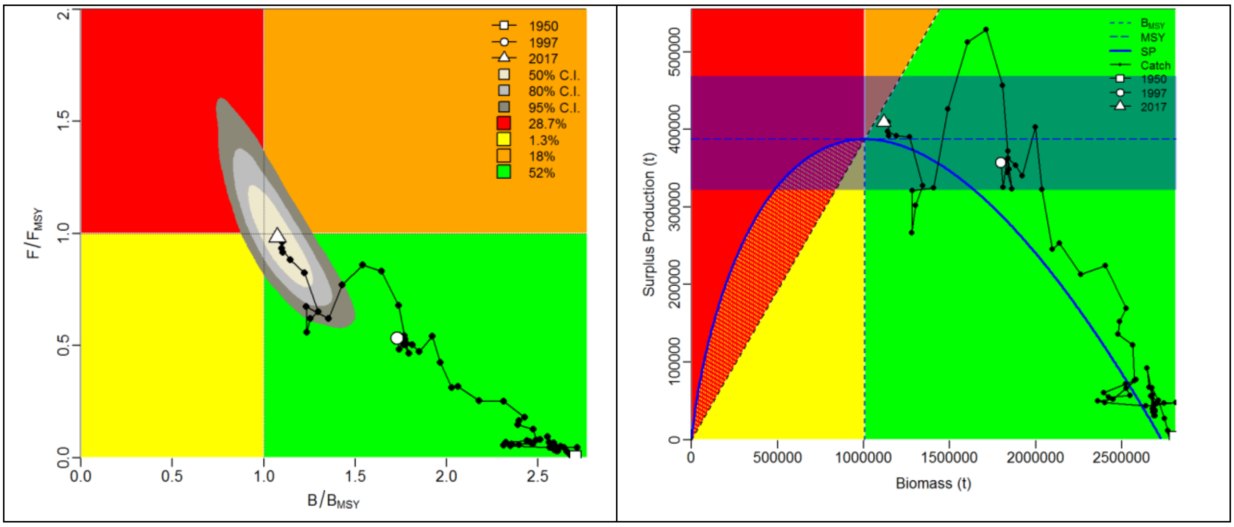

Figure 3: Kobe plot and MSY and Kobe dynamics

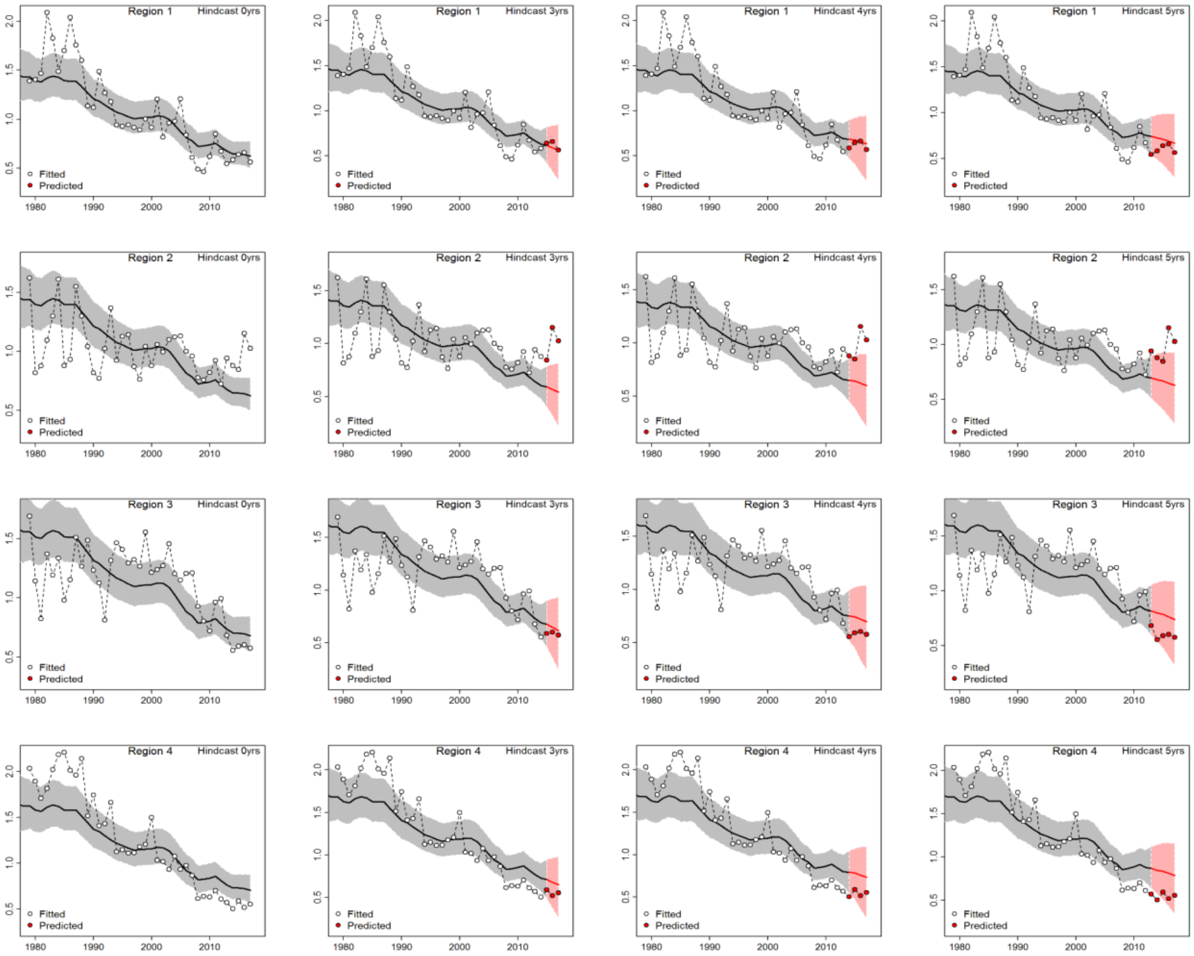

Figure 4: Hindcasting on different areas and the JABBA model for 3, 4 and 5 years.

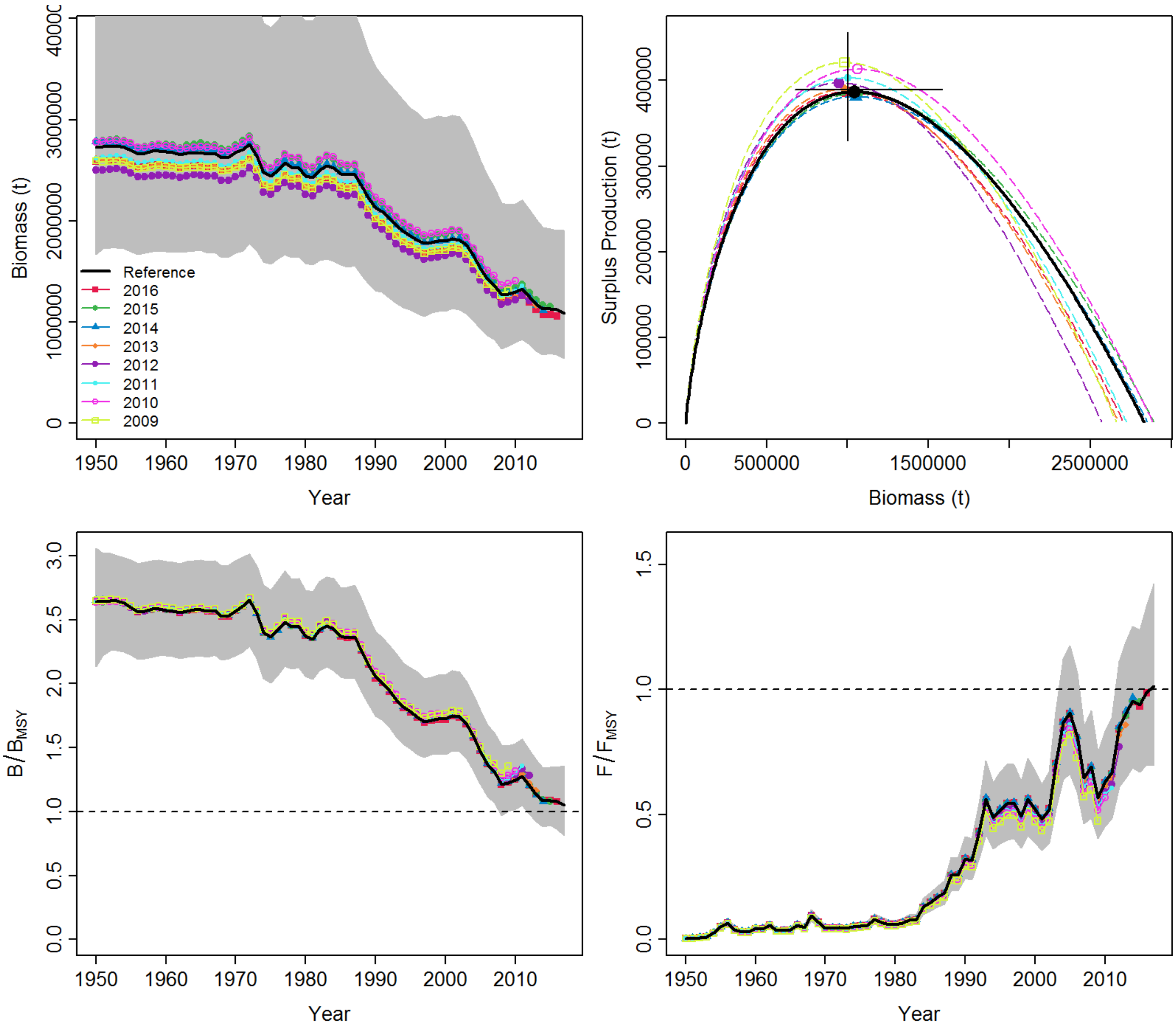

Figure 5: Retrospective analysis for JABBA model from 1to 8 years

In addition hindcasting indicates that the model does fairly well in predicting trends in Areas 1, 3 and 4. However, Area 2 has poor predictive power with the generated credible intervals shown. Retrospective patterns (Figure 5) are not as pronounced as the SS assessment which constantly appears to be under predicting.

iii) Integrated Assessments (SS3).

Background Material/Model Specifications: Stock resolution 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. 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 there 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 6&7 from Langley et. al. 2012).

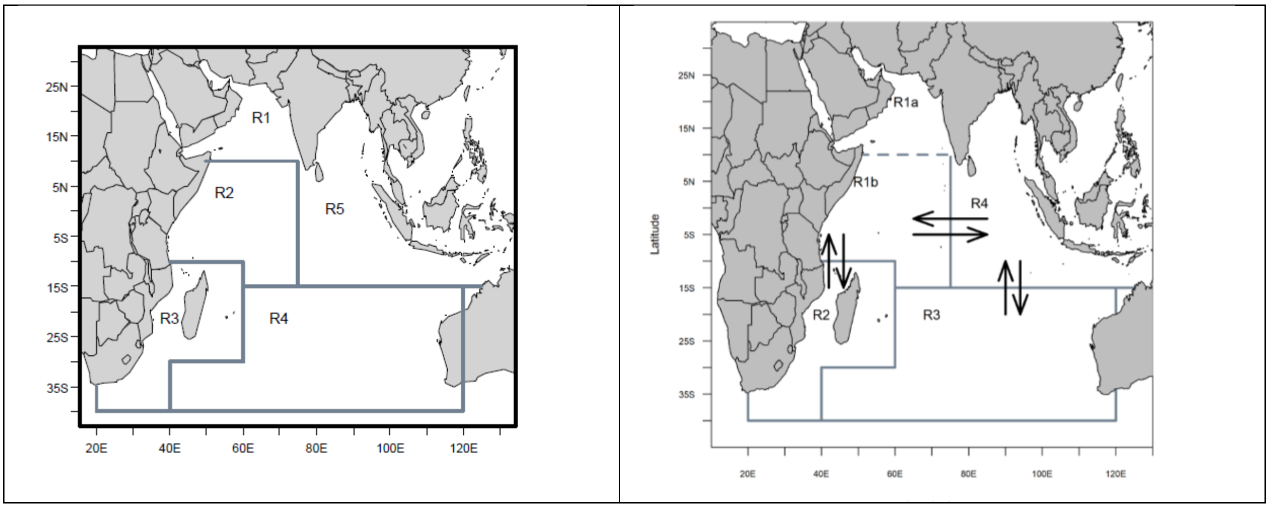

Figure 6: Spatial coverage from 2012 vs 2018

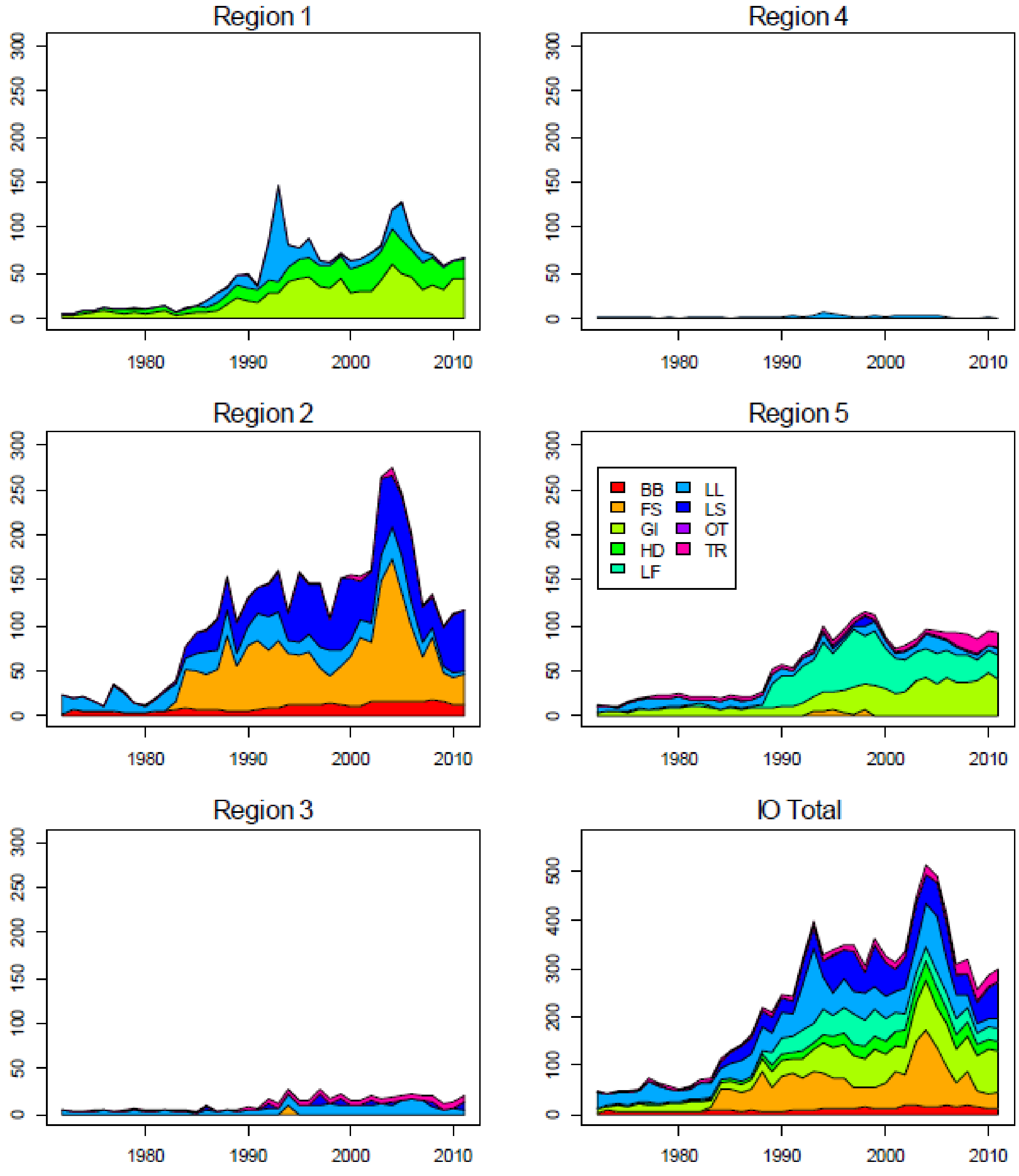

Figure 7: Catch Distributions by 5 areas (from Langley et. al. 2012).

Iterative reweighting approaches should be examined using some justification for weights (likelihood profiles of what influence the assessment should be examined). As such, a recommendation to further down-weight 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 LF weights.

The main issues identified in the original runs presented were:

1. LF data after 2003 is suspect, so either down weight or ignore it (suggestions to exclude data after 2013, as it was primary reason for retrospective pattern). However this still means the LF data from 2003-2013 are providing problematic data to fit to.
2. Examine in more detail the effectiveness of tags on the assessments in conjunction with other datasets in the assessments.
3. Profile likelihood plots on R0 are useful diagnostics for assessments, as this may indicates what may be driving the assessment and it would be clear whether the assessment was being influenced by the LF, or the CPUE or tags. If the latter, (see Francis 2011) one should fit to indices of abundance rather than LF.
4. In addition to R0 profiles, jitters to assess whether there were global minima, and retrospective patterns should be examined. While the latter was examined, the former was missing.
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.
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 to use are. The grid structure really was not examined in detail to assess these effects.
7. Uncertainty needs to be accounted for as much as possible, as currently it mayb e underestimated. Possible approaches are to use grid based versus MCMC based approaches, though computation time makes this difficult. I would recommend using main effects analysis on a partial grid to assess how things may change over time, and also use in projections, as it would save time. A more comprehensive examination of uncertainty using MCMC at a later point on a reference case model should be conducted.

Other issues and possible solutions:

1) How do we deal with recruitment (continuous or recruitment at a particular time)? Examining recruitment trends are important as they may be the main factors affecting the assessment, and need to be accounted in some context. Issues related to recruitment only occurring in area 1 and 4, can be reexamined and dispersed in all areas as its normally done.
2) How to address correlated parameters over time? How do these affect the assessment?

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.

It is also **HIGHLY RECOMMENDED** to do a jitter analysis to test for convergence and global versus local minima issues.

OTHER CRITICAL ISSUES THAT WERE NOT ADDRESSED:

a) **Alternative CPUEs :** In essence, we can't believe two conflicting indices at the same time. Either we believe the LL index or the PS index (but not both in the same area). Hence, it would be recommended to use one index for any given area. My recommendation would be to fit alternate indices separately as these may drive the assessments, and test alternative hypothesis. In addition fitting to different effort creep assumptions for the PS fleets are also useful, to understand how these assumptions influence these assessments. Also examining whether 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 Kitikado at the meeting. Examining different 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 (find some more) 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.

b) **Data weighting and conflicting sources of information for assessments:** Based on the conflicts in length frequency and index of abundance in the datasets, using some down-weighting of the LF data or ignoring it entirely is recommended. 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 composition data, allow for correlations; and Principle 3: Do not down-weight abundance data because they may be unrepresentative." 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

c) **Dealing with Uncertainty:** In addition, when using forecasts, MCMC based projections could be examined. These are important analyses that need to be addressed in these assessments which were not accounted for at this meeting. It is not clear how we deal with projected catch advice as we currently may have an overfished stock (based on reference cases presented in 2018WPTT 20-33).

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 and stock status advice should be examined.

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, 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. If of any use, the JABBA based estimate of K (2731157) vs S_{B0} from SS (2,779,850) vs yield targets, MSY is 388Kt (JABBA) vs 377Kt (SS3 from reference case in WPTT 20-33)

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

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 3 years some considerations should be given to this, unfortunately no discussion or time was spent on this.

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.***

Initial runs did not show any information on likelihood values of the fits and different components. In addition, profile likelihood approaches should be used for different components and how the effect the profiles (i.e. length composition, tagging, and survey/cpue data). These should have probably been done on other parameters (R_0, and MSY for example) and will also cover issues of non-convergence or other issues while parameterizing the models. Structural uncertainty was not examined due to time constraints; however data-weighting or alternative area examinations (5 versus 4 areas in the integrated assessment for example) 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, 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 (however due to large run times the 1st was discounted, and the 2nd was not examined either).

The final "sensitivity (not grid)" runs decided on were determined on a very arbitrary basis. It appears that in the one-off sensitivity analyses, not much change was observed by adding additional CPUE's or weighting the tagging data and down-weighting the LC data. The results were surprising giving the value of information obtained from the tagging data, and the interactions with growth, natural mortality, selectivity, and the length data. While issues were discussed on effort creep scenarios and other factors, the final runs did not include a lot of these scenarios or interactions, neither was growth uncertainty incorporated in the assessment. 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.

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 hypothesis 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.

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

Not relevant as MSE in development currently.

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. 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. However, some papers such as the Pole and line index were written in Markdown, so that all analyses are transparent and replicable. Such methods would be useful as a basis in future years for the assessment models and possibly other analysis that is presented at these meetings.

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 a lot of time was spent discussing alternative model runs and approaches, as well asthe data at the meeting, I suggest the following steps to streamline the process:

a) All datasets are made available to the modellers 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). <u>**More time spent at the data meetings clearing the data issues of developing coastal**</u> <u>**countries that have important fisheries on the species that is the**</u>

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.

This is not an overly critical review of the approach, but 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.

Finally, while, a reference model is good for advice, numerous alternative models should be examined, and possible a grid of models should be presented to show uncertainty in the assessment and for projections, as a lot of the biological parameters are not known well in these assessments**.**

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. 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), 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. 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. In addition, examination of uncertainty using MCMC/bootstrap approach on all models is important to assess the adequacy of the sensitivity runs. 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

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.
2) 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.
3) 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 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.
4) Issues of initial tag mortality (Hoyle et. al. 2015) really need to be directing the work here. It seems contrary to normal thought that a peer reviewed paper would not be used to the base case. I strongly recommend using the initial tag mortality rates that Hoyle (2015) state, i.e. it should be 20% as the study suggests and not 2-3% that Gaertner and Hallier (2015) study suggests as these were designed and analysed for different objectives.
5) 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). Paper 42 (IOTC 2018-WPTT 20-42) indicates that the previous model in 2016 had issues with local minima, and also had retrospective patterns which were a cause of concern. Although some checks were done at the meeting, insufficient time was spent on diagnostics that need to be accounted for at a later period.
6) Issues of spatial complexity; going back to 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. However, alternative hypothesis of catchability drop for LL fleets would give a very different outlook on the stock, and 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 good way to go. 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 (particularly for PS fleets).

Overall Conclusions

The use of multiple approaches is important when assessing stock status. While different approaches were examined (ASPM vs SS), enough time was not spent on diagnostics (jitters, profile likelihood, and retrospective analysis), 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 on what the final models to use for advice were developed without a thorough analysis. **A major issue was an issue with local and global minima that was not investigated thoroughly (Jitters, and retrospective analysis for example were not investigated thoroughly).** Length frequency data are particularly important for SS, and as such examining if these data are accurate is critical in the assessment. 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 needs 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 more substantial analysis should be conducted on this issue.

Overall, the process was transparent, and issues were discussed (although maybe not extensively). 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, 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).

Acknowledgements

I would like to thank the secretariat (namely Dr. Chris OBrien, Dr. Paul DeBryun, Dr. Fabio Fiorellato, Dr. Sarah Martin and Mr. James Geehan) for giving me the opportunity to come as invited expert for WPTT 2018. I would also like to thank the Chair, Dr. Shiham Adam, vice chair Dr. Gorka Merino, and SC chair Dr. Hilario Murua 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