REPORT OF THE 2020 ICCAT INTERSESSIONAL MEETING OF THE BLUEFIN TUNA SPECIES GROUP

(Online, 14-22 May 2020)

The results, conclusions and recommendations contained in this Report only reflect the view of the Bluefin tuna Species Group. Therefore, these should be considered preliminary until the SCRS adopts them at its annual Plenary meeting and the Commission revise them at its Annual meeting. Accordingly, ICCAT reserves the right to comment, object and endorse this Report, until it is finally adopted by the Commission.

1. Opening, adoption of agenda and meeting arrangements and assignment of rapporteurs

The online intersessional Bluefin tuna Species Group ("the Group") meeting was held from 14 to 22 May 2020. Drs John Walter (USA) and Ana Gordoa (EU-Spain), the Rapporteurs for the western Atlantic and eastern Atlantic and Mediterranean stocks, respectively, opened the meeting and served as Co-Chairs.

The SCRS chair initially addressed the Group, noting the challenging circumstances under which the meeting was being held. As matters progressed during the meeting, several themes emerged related to these challenges and to the path forward. While this text was originally presented to the group at the end of the meeting, we have placed it here as it provides context to the remainder of the report:

"The COVID-19 pandemic has had a significant impact on the standard operating procedures of the SCRS and the ICCAT Secretariat. National and international restrictions to combat the crisis has resulted in the cancellation of some scheduled meetings, the rescheduling to later in the year of others, and even the postponement of some until 2021. A major outcome of the restrictions has been the inability to hold face-to-face intersessional and working group meetings. Internet/webinar meetings have become the new way of doing business. Because of this, in March of 2020, an initial decision was made that the SCRS would primarily focus on and restrict its activities to the proposed stock assessments necessary to provide advice to the Commission. Subsequently, as time progressed and CPCs became more comfortable with on-line meetings, it was discovered that this format was a partly viable solution/alternative to the traditional meeting, at least for some topics. However, it is important to note that webinars cannot replace face-to-face meetings for transparency and some decision-making processes that require CPC representation, broad discussion input, and a necessary consensus to proceed. Unfortunately, from a practical perspective, webinars may be the only way of doing business for some time to come and major decisions may have to be made at on-line meetings. The SCRS and the Secretariat are exploring ways to achieve an inclusive on-line decision-making process, but we at not there yet. Currently many of these decisions are being postponed until the next face-to-face meeting.

For MSE, the path forward described in this report represents an aspirational proposal to conclude the MSE work in time for the adoption of an MP in the Autumn of 2021. Unfortunately, due to the current pandemic that has precluded the in-person meetings originally deemed necessary to complete tasks such as the adoption and plausibility weighting of the reference grid, the MSE process will probably, unless an alternative can be found, experience a delay until these and other critical tasks can either be completed through in-person meetings or arrangements are made to take decisions through remote meetings. At the moment, we are not sure that we can or want to take challenging decisions online, but we may be forced to if the crises extend beyond the short term. Utilization of remote meetings will assist in the completion of many of these tasks and presumably a mutually agreeable mechanism will be available for decision-making, if necessary, in the long term."

On behalf of the Executive Secretary, the Assistant Executive Secretary welcomed the participants to the meeting. The Chairs proceeded to review the Agenda which was adopted after some changes (**Appendix 1**).

The List of Participants is included in **Appendix 2**. The List of Documents presented at the meeting is attached as **Appendix 3**. The abstracts of all SCRS documents and presentations provided at the meeting are included in **Appendix 4**. The following served as rapporteurs:

Sections	Rapporteur
Items 1, 12	A. Kimoto, N.G. Taylor, M. Ortiz
Item 2	H. Arrizabalaga
Item 3	M. Ortiz
Items 4, 6	A. Kimoto, N. Duprey, K. Gillespie, E. Andonegi
Items 5, 7	C. Fernandez
Item 8	A. Pagá
Item 9	G. Melvin
Items 10, 11	J. Walter, A. Gordoa

2. Review of the scientific papers relevant to assessment

Three documents were presented under this agenda item.

SCRS/2020/063 updated the Gulf of St. Lawrence (GSL) acoustic index. This index has been consistent with fishery CPUE values from the Gulf of St. Lawrence for most of the time series, but recent updates (2017-2018) suggest a significant decline in bluefin tuna (BFT) that does not appear reflected in fishery catch rates. The authors investigated potential factors that may explain the recent lower index values (changes in data editor, vessel and abundance of herring). Results suggest that the recent low index values do not appear to be related to the data editor, vessel or to trends in herring biomass; however, there does appear to be a requirement for a minimum biomass of herring. Thus, the authors suggested that index values significantly below this minimum could be removed from the index.

The Group noted that in the historical time series, there were several years where herring density was below the suggested threshold (0.25 kg/m^2), but BFT abundance did not seem to be reduced that much. The Group also noted that while the argument that BFT presence might be driven by prey, this would probably not only have an effect when herring abundance is low, but also when herring abundance is high. The Group agreed that critically low food levels might drive BFT out of the Gulf, but in general this was not necessarily a binary in/out problem, but something that needed to be further studied in the future to see how best to standardize the index regarding prey abundance. The Group also noted that these efforts might be limited by the fact that the surveyed area and survey duration were relatively limited and might not represent the BFT spatio-temporal distribution within the Gulf, regardless of whether the index value was high or low. Further investigation of the potential impacts of oceanography on predator or prey distribution coupled with incorporation of the large body of electronically and acoustically tagged fish in this region would be fruitful avenues of research to assist in the interpretation of this index.

Beyond index standardization, the Group discussed alternative approaches to deal with this issue within the assessment model, such as considering time-varying and environmental-dependent catchability. The Group was reminded that within Stock Synthesis an environmental factor was used to modulate availability as a function of oceanography, which affects not only this but also other indices, and suggested further exploration of this point. Alternative suggestions were to consider variance inflation factors that could accommodate processes such as availability to the Gulf of St. Lawrence. This could be an option for the future, but not for the strict update to be conducted in 2020, as the CVs are fixed and no additive variance is considered for fishery-independent indices.

Based on this discussion, the Group decided that the base-case run of the assessment would include the acoustic index but not the 2018 data point, and they would conduct a sensitivity analysis including the 2018 data point. The sensitivity run would be to check the impact of having dropped that point and could be used to provide qualitative advice; it would not be used in projections or for generating Kobe matrices. The authors of the document informed the Group that the 2019 data point is expected to be available before September, so might also be useful for qualitative advice.

SCRS/2020/067 analysed how the Balearic larval index is affected by different configurations for the same environmental variable (temperature in the mixed layer depth) and different modelling approaches (nonlinear Delta-log, delta-gamma, tweedy and Bayesian). The authors also investigated the effects of differences in total sampled area among years on the index. The results were used to interpolate larval index values in years with no standard larval surveys but with some ichthyoplankton surveys available, and to propose a "revised version" of the index and its uncertainty.

The Group agreed to further discuss potentially replacing the currently used index with this new proposed index in the future. The Group also mentioned that different age classes might have different fecundities, and this might affect the variability of the index. This is something that might be explored in the future, together with other issues like survival due to food and temperature. Finally, the Group also asked whether 2018 and 2019 data points will be available before September or not. The authors responded that the 2019 data point will not be available, and the 2018 data point is also unlikely to be available.

SCRS/2020/068 described an age calibration exercise. The Fish Ageing Services laboratory (FAS) was contracted by ICCAT GBYP in Phase 7 to provide age estimates from 2000 Atlantic bluefin tuna otolith samples. With the objective of ensuring that age readings provided by FAS follow the ICCAT reviewed reading protocol, a sub-sample calibration exercise was carried out. Band counts are similar between FAS and the group of laboratories involved in direct ageing. However, there is a one-year bias in the count of bands in older specimens, starting from 10-13 years of age, with a lower count by FAS compared to the rest of the laboratories. This bias seems to be due to the fact that FAS counts the bands in a different area of the ventral arm of the otolith. It would be necessary for FAS to reread the samples of specimens older than 10 years using the area close to the sulcus margin of the ventral arm. Age quality control with these and other otoliths would allow for several thousand age readings be available for the next assessment.

3. Presentation of initial data inputs

The modelers presented a summary of the data inputs for the updated stock assessment models for East (SCRS/2020/069) and West BFT stocks (SCRS/2020/070 and SCRS/2020/072). The Chair emphasized that according to the work plan approved in 2019 (Anon., 2019c, Appendix 5), the assessment follows as much as possible the concept of a "strict update" of the 2017 models used to provide the advice. To be consistent with the "strict update" concept, input data should be as similar as possible to those used in 2017.

3.1 Biology and age data

With regards to biological data, the modelers will use the same parameters and assumptions as in the 2017 stock assessment (Anon., 2017b). It was noted that since 2017, a considerable number of otoliths and age readings have been made available to scientists for the West stock, in particular for earlier years (2015), and the Group welcomed this new information. Details on how the age data were treated in stock synthesis are provided in SCRS/2020/072.

3.2 Size and age composition

The Secretariat provided the CAS (catch-at-size) and CAA (catch-at-age) for East and West BFT stocks, as well the size frequency samples for the western BFT Stock Synthesis fleet structure input incorporating data submitted by CPCs before the deadline of 5 April 2020. The CAS was updated only for the years 2016-18, incorporating new information from the Task2SZ (size data), CAS submitted by CPCs, and the stereo-camera size data from caging operations. The overall CAS construction used the same substitution rules as in the 2017 stock assessment (Anon., 2017b), with one exception: for the fleet Mediterranean OTHERS purse seine (Anon., 2017a, Table 3) the size distributions from stereo cameras in 2017 and 2018 were excluded as their size distribution patterns varied substantially from prior years (2015/16 and 2019). Unfortunately, as there was not enough time to conduct a detailed review of the submitted data, size density distributions for 2015/2016 from the same fleet were used to estimate the CAS for 2017/18, respectively.

Further review of the CAS and size data revealed other inconsistencies that were addressed for this assessment update. Briefly these changes include the following: a) removing from the PS Croatia size samples of fish harvested at the end of farming operations, and b) conversion to SFL (straight-folk-length) for US longline size samples (1996, 2000-2010) that were originally reported as curved pectoral fork-length. In addition, it was noted that the Secretariat adjusted CAS submitted by Japan (2009-2018) to match Task 1 NC, however their national scientist indicated that this adjustment was not required as they census their full catch. Nevertheless, as Task 1 NC and CAS should represent the same biomass, the Secretariat requested that details be provided on the size-weight conversion factors used to estimated CAS, so they can be compared to the official Task 1 NC reported. Finally, the Group recommend that the Secretariat work in collaboration with national scientists to carefully review the Task 2 stereo-camera size data submitted by the fleet Mediterranean OTHERS purse seine for 2017-18 and confirm the correct size distribution of their catch.

The CAA for both East and West BFT was estimated using the ageit software, the same slicing algorithm used in 2017. The age is assigned by inverting the growth function by month, in the case of the East BFT the von Bertalanffy growth model (Cort *et al*., 1991) with parameters (Linf = 318.85, k = 0.093, t0 = -0.97), and for the W-BFT the Richards growth model (Ailloud *et al*., 2017) with parameters (Linf = 270.6, La = 33.0, Aa = 0.0, Ab = 34.0, p = -0.12, k = 0.22). The comparisons with the CAA inputs for the 2017 stock assessment (Anon., 2017b) showed very similar age distributions for both East and West stocks. Partial CAA were also estimated for both stocks with a similar age composition as in the 2017 stock assessment.

3.3 Catch Estimates

3.3.1 Task I Nominal Catches

The Task 1 Nominal Catches (NC) time series provided by the Secretariat for this assessment matches almost exactly with the catch series used in the 2017 stock assessment (Anon., 2017b) for both East and West stocks. There were minor differences in 1995 due to review of the Spanish baitboat historical series, the regular updates for the last year in the 2017 stock assessment (2015), and the new catch information for 2016-18. The Group noted that, whenever the historical data sets coming from the so-called "Brazilian catches" (large catches of BFT in the South Atlantic between 1960 – 1970) are used in an assessment model, as it was done in the past for the West BFT or currently in the western Stock Synthesis, the discussions occurring in the February 2020 BFT MSE Technical Group (BFT-MSETG; Anon., 2020b, *in press*) on the stock-of-origin of this specific data set should be always considered, given that the origin of these important catches is still unknown and undefined.

3.3.2 Catches 2019 and 2020 assumptions for projections

It was noted that for 2019 catches, the TAC (32,240 t East and 2,350 t West) will be assumed for the assessment projections, as no other information indicate the contrary. However, while 2020 catches will also be assumed to match the TAC (36,000 t East and 2,350 t West) the COVID-19 pandemic could impact on this year's realized catches, so there is a possibility this will need to be reconsidered later on for stock projections. The Secretariat will provide updates on 2020 catches at the second intersessional BFT Species Group (BFTSG) meeting in July and again in September's BFTSG meeting.

A comment was also raised in regard to the escapement of bluefin tuna from Spanish Mediterranean farms due to recent storms in the area, indicating that part of this biomass may return to the wild population. However, there is no information on the total biomass lost or the mortality of these fish. Therefore, the Group decided not to deduct the amount of presumed escaped fish from the catch.

3.3.3 IUU catches

Possible IUU (Illegal, Unreported and Unregulated) catches in the Mediterranean are under investigation and, until more information is available on the magnitude and timing of these catches, the Group cannot speculate on their impact on the assessment or assessment advice.

Neither the Secretariat nor the SCRS has received new information regarding the ongoing investigation of Mediterranean IUU catches of BFT. Therefore, the Group does not have a strong basis for estimating their magnitude or to include them explicitly in the current assessment. As with any IUU, the absence of information on trend and magnitude poses a challenge to the integrity of the advice that the SCRS provides to the Commission, advice which currently does not assume any IUU. Hence the Group requests that all available information on this situation be forwarded to the Secretariat as soon as possible.

3.4 Indices of abundance

Indices of abundance updates and changes were reviewed during the SCRS bluefin tuna Species Group meeting in September 2018 (Anon., 2019d). The Canadian acoustic survey index and the Western Mediterranean larval survey index (W-Med Larval index) were provided by authors on time by the end of January, and the Group agreed to use them in the current assessment. The compiled table of available indices is shown in **Tables 1** and **2**.

As indicated in Section 2.1, for the GSL acoustic index it was decided to exclude the 2018 data point in the base model and include the full time series as a sensitivity run. The Group had extensive discussions of a number of factors that affected the index including a change of vessel, which was the rationale given for excluding the index from projections in the MSE (February BFT-MSETG meeting, Anon., 2020b, *in press*) the relationship between herring biomass and probability of detection and the potential for changing spatial variability in the availability of bluefin tuna to the survey. The Group excluded the 2018 data point on the basis that this year was likely an anomaly due to changing spatial distribution of BFT, either as a result of changes in local or overall herring abundance or oceanographic conditions and the fact that the survey covers only a small area relative to the entire Gulf of St Lawrence. Such a conclusion was corroborated by the high catch rates in the fishery dependent index indicating no similar change. Further work evaluating the spatial availability of bluefin tuna to the sampling frame of the acoustic index relative to prey abundance and oceanographic conditions would greatly benefit future evaluation of this index. Additionally, evaluation of the extensive electronic and acoustic tagging information within the Gulf of St Lawrence should further elucidate whether 2018 was an anomalous year for spatial distribution of bluefin tuna in this region. It should be noted that the Stock Synthesis model explicitly assumes that the catchabilities for this index, the GSL-SWNS index, and the USRR>177 index are all affected by the Atlantic Multidecadal Oscillation (AMO), hence implicitly assuming oceanographic influences on availability.

4. Review of initial assessment diagnostics

4.1 East (VPA)

SCRS/2020/069 and SCRS/P/2020/020 presented the data and model set-up for the 2020 eastern and Mediterranean Bluefin tuna stock update of the 2017 stock assessment (Anon., 2017b). For the present analysis, the data over the historical period (1968-2015) were nearly identical, whereas the data for the years 2016-2018 and abundance indices were updated. As agreed in previous meetings, initial model specifications were kept identical to the 2017 assessment.

The paper had three different sets of runs for comparison: a 2017 base case (Run0); the same model specifications as Run0 with the updated data from 1968-2015 (Run83); and, the same model specifications as Run0 with the updated data from 1968-2018 (Run84). The Group discussed that Run0 and Run83 had very similar trends in SSB (spawning stock biomass). This was also the case with the recruitment trends until the last year included in model estimates in 2012. The last three years of VPA recruitment estimates were discarded as they are poorly estimated.

The Group had extensive discussions on the SSB, recruits (age 1) and F-ratio (F_{age10+}/F_{age9}) trends, especially on the Run84 results. For Run84, the rapid increase in biomass in the later years of the time series appears to be very strong. Also, there are very different trends in the number of recruits between Run0/Run83 and Run84 after 2007, with the number of recruits being much higher in Run84. The Group highlighted the severe retrospective bias in that model and multiple solution states. Further, the Group discussed that the last year of data has noticeable effects on the trends in SSB and recruits (**Figure 1**). The Group also noted that the jackknife plots showed that the removal of some indices heavily influenced SSB and the number of recruits (**Figure 2**). In both the retrospective and jackknife plots, the F-ratio showed large fluctuations between treatments from the late 1990s to late 2000s.

The Group discussed how the 2017 assessment (Anon., 2017b) also had a significant retrospective pattern and F-ratio issues. Since only new data is being added and no model specifications are changing, these problems should be expected to persist. The Group also discussed an issue with the diagnostic that examined the sensitivity of results to the jitter of starting values. It was evident that the MLE (maximum likelihood estimate) solution for SSB and recruitment etc. did not fall within the envelope of the curves defined by the jittered values. The Group discussed if the MLE estimates represented a biologically plausible solution (extremely low *Fs* compared to the 2017 model and alternative solution state to other jitter runs), and decided the model required further examination.

The Group agreed to have several new sensitivity runs prepared by the analysts for review later in the meeting, with the aim of evaluating alternative models based on diagnostic criteria improvements. Primarily, model selection was based on 1) consistent model convergence to a global solution at different starting parameters and 2) the level of retrospective bias observed when removing one to five years of data.

The three alternative runs were: Run84 with F-ratio fixed to the values estimated by the 2017 base case; Run84 with the F-ratio fixed to the values estimated by the 2017 base case and then estimated for 2016-2018 as one time block; and, Run84 with the W-Med Larval index split into two (period 1 = 2001-2005 and period 2 = 2012-2017 with a separate catchability estimated for each time period).

SCRS/P/2020/026 presented the alternative VPA runs prepared by the analysts that were requested earlier in the meeting. In Run133 (which is Run84 revised to set the F-ratio fixed to the values estimated by the 2017 base case, **Table 3**), the retrospectives patterns were improved, but stability issues continued with the jitter of starting parameters. Run135 (Run133 with a change to the lower bounds for estimates of terminal *F*) improved model convergence by fixing the F-ratio to the 2017 values and changing the lower bounds for estimates of terminal *F* (0.005, 0.02, 0.02 and 0.01 instead of 0.1e-6 for ages 3-6, respectively) (**Figure 3**). Run135 jackknife indicated that SSB and recruits estimates are less sensitive to individual indices (**Figure 4**). While some retrospective patterns remained in Run135, **Figure 5** clearly shows a large improvement in the SSB retrospectives. However, recruitment estimates from 2008 onwards varied drastically in scale when peeling off individual years but were less variable than in Run133. The Group discussed the high uncertainty in recent recruitment estimates as a major limitation of the model, particularly regarding the consequences on projections.

Run174 matched Run135, with exception of the estimated F-ratio for the terminal period (2016-2018). This run showed similar results to Run135 sensitivity runs in most aspects but with improved retrospective bias metrics. However, Run174 did not fix the overall retrospective pattern in recruits of high magnitude change across retrospective runs (**Figure 5**). The Group concluded that Run135 (fixing F-ratio to 2017 values and adding a change to the lower bounds for estimates of terminal *F*) would be used as there was no significant statistical improvement in the objective function value obtained by adding the extra parameter, nor substantial evidence of its estimability from likelihood profiling of this parameter.

The Group also reviewed the runs where the W-Med Larval index was split into two blocks of time (splitting period 1 = 2001-2005 and period 2 = 2012-2017 with a separate catchability estimated for each period). This resulted in Run117 (Run84 with the W-Med Larval index split), Run185 (Run135 with the W-Med Larval index split), and Run186 (Run174 with the W-Med Larval index split) (**Figure 6**). There was no improvement in statistical likelihood (addition of one catchability parameter) or retrospective bias as a result of splitting the index, apparent across the larval index sensitivities (see **Figures 5** and **6**). Considering the lack of improvement in the model information content or retrospective pattern — the key reason for evaluating the split — the Group selected Run135 as a preferred base model.

4.2 West (VPA and Stock Synthesis)

West (VPA)

SCRS/2020/070 provided the work performed regarding the update of the indices used in the current assessment model. Comparisons between the indices used in the 2017 assessment (Anon., 2017b) were shown, particularly emphasizing the error found in the filter for the Japanese Longline partial catch-at-age data. This bug has already been fixed for the 2020 assessment model, causing little differences overall.

The first diagnostics were not shown in this document, mainly due to the Group's decision to remove the 2018 Gulf of St. Lawrence Acoustic index value from the base model. In addition, it was necessary to change the input CVs of the GOM (Gulf of Mexico) larval index to obtain a minimum fishery-independent indices value of 0.3 and not estimate additional variance terms for those two indices (GSL Acoustic and GOM Larval survey). All the analysis will be shown in the next assessment session on Monday, including the Jackknife, jitter, age-plus-group F-ratio, retrospective, and sensitivity analysis, the latter focusing on the sensitivity of the current model to the use of the Canadian Acoustic Survey Index in 2018.

SCRS/P/2020/027 West BFT VPA results were presented for the revised base case during the meeting, showing relatively stable recruitment estimates, which suggest that there is good convergence of the models (**Figure 7**). Similar recruitment estimates are obtained from all jitter analyses (**Figure 8**). The model fit to indices was shown, with no strong patterns on the residuals of these indices, except the trends found in the JPN_LL index residuals. Regarding the Jackknife index analysis, it was shown that removing the GOM-Larval index causes higher divergences, especially between the SSB levels (**Figure 9**). However, across all index exclusion sensitivities, the stock biomass (younger spawn scenario) was predicted to decline for a period of four years.

The retrospective analysis (**Figure 10**) showed some divergences in both recent recruitment and spawning stock biomass (in the younger spawning fraction scenario). Furthermore, the analysis of the F-ratio profile suggested that alternative assumptions of an F-ratio higher than 1 were supported by the model information, differing from the conclusions of previous assessment. The Group noted that differences in the F-ratio could be influential on model results. The bootstrap analysis of fishing mortalities showed a very low F for ages 1 to 3, with higher and increasing values for ages 8, 9, 10 and 11.

The use of CAN_RR and US_RR_177 in the Jackknife analysis for recruitment was raised, but the authors clarified that it was not included in the models and therefore it was not included in the figure since it has a 0 in front of the label, meaning that it is not modeled in the VPA.

Regarding the retrospective analysis, the issue of divergence between the results in 2017 and in 2020 going back 3 years in the retrospective analysis was raised. Authors explained that this divergence could be explained by two main reasons: 1) because the indices were for small fish and had high leverage on recent recruitment); and 2) because the search for the minimum log-likelihood is not conducted for each retrospective peel. This search is conducted through a time-intensive jittering of the starting values and is critical to obtain overall convergence for advice models but is not usually conducted for individual peels. Had it been performed, estimates may have been more similar in terms of scale. There is a slight divergence in recruitment estimates but, overall, the two 2017 models and the 3-year retrospective peels are quite similar (**Figure 10**).

The question of how historical Brazilian catches from the 50s should be treated in the assessment models was highlighted. As this is an issue related to the basic data, it is appropriately taken up under Section 3. While the issue of the stock of origin of these fish remains uncertain and is a robustness test in the MSE, the Group noted that a strict update would not alter decisions at the 2017 assessment (Anon., 2017b) regarding the basic treatment of this data across the various modeling platforms (East/West and VPA and Stock Synthesis).

The analysts confirmed that differences in the F-ratio affect the assessment's scale (SCRS/2020/070). However, this was less of a problem in the West assessment compared to the East stock assessment, because the West stock uses a plus group starting at age 16, whereas the East stock plus group starts much younger at age 10. As a result, differences in the assumed F-ratio (F_{age16}/F_{age15}) are less influential. Furthermore, the Group raised the point that issues regarding the implications of the F-ratio (synonymous with dome vs flat selectivity) could more easily be explored with Stock Synthesis where the assumptions regarding dome-shaped selectivity on older ages could be explored on a fleet-specific basis by looking at composition data and estimated *Fs* at age.

West (Stock Synthesis)

SCRS/2020/072 described initial model set up, fleet definitions, selectivities and parameterizations for the 2020 western Atlantic Stock Synthesis (version 3.30) model. Input and model settings underwent only slight changes from those used in 2017, commensurate with the strict update. Two model configurations (early and late maturity) span from 1950 to 2018 and were fit to length composition data, conditional length at age data, 13 indices and 13 fishing fleets. One CPUE index (US GOM LL) was accidentally split in the wrong year for the 2017 assessment (Anon., 2017b) (1987-1992, 1993-2015). The correct split date was implemented for this assessment (1987-1991 and 1992-2018).

The Group agreed that the base model runs need to be updated to omit the 2018 Canadian Gulf of St. Lawrence acoustic index data point (see Section 2). The Group also decided to include the 2018 data point as a sensitivity run.

SCRS/P/2020/023 provided the summary of diagnostics results (jitters (**Figure 11**), retrospective (**Figure 12**), recruitment likelihood profiles (**Figure 13**), Jackknife (**Figure 14**), parametric bootstraps (**Figure 15**)) of the revised base-case model for the 2020 western BFT Stock Synthesis model with the removal of the 2018 Canadian acoustic survey data point.

Model settings remained largely unchanged from the 2017 assessment (Anon., 2017b). The authors noted that diagnostic results showed similar patterns to those observed in the 2017 assessment. There were minor issues regarding fit to data and some of the diagnostics results; however, these were also present in the 2017 assessment.

The Group noted that the log-likelihood profile (**Figure 13**) for virgin recruitment shows that the large influence of the recruitment deviations from the assumed Beverton-Holt stock recruitment relationship.

The authors further noted that, as in 2017, sigma R was estimated using the bias correction ramping, an ad-hoc tuning procedure that acts with prior-like behavior. It was noted that the MLE sigma R was different from the sigma R observed in the bootstrap distribution. It was clarified that this is likely due to the bootstrapped datasets-length composition data were resampled, and the subsequent data were probably not as variable as the real data, resulting in likelihoods that were better than those from the base model.

Given the scope allowed for this strict model update, the Group noted that further tuning of the model was not needed.

SCRS/2020/071 examined the relationship between the Atlantic Multidecadal Oscillation (AMO) and indices of abundance for the northwest Atlantic. The authors re-examined correlations between indices and the AMO and presented an updated sensitivity run for the 2020 West BFT Stock Synthesis assessment incorporating the AMO. Results indicate that the AMO continues to be correlated with U.S. and Canadian abundance indices, and incorporating the AMO in the assessment model does not substantially alter parameter estimates from base scenarios but does improve model fit to U.S. and Canadian indices.

The Group noted that the inclusion of environmental parameters in assessment models presents a good method for investigating biological hypotheses about the stock. They noted that the inclusion of additional covariates should always be treated with caution, as in many cases the inclusion of these additional covariates can result in better fit regardless of their relationship to the response variable. It was further noted that inclusion of additional effects can serve to flatten or obscure abundance trends. The Group suggested that clear indicators be used to assess the assumption that the AMO was not simply improving fit but was also improving the overall model. The authors noted that this is an active area of investigation and the further analyses will be presented to the Group.

The Group asked for clarification on how AMO patterns were implemented within the model. The authors explained that the AMO index was used as a time-varying parameter that modulates catchability within two fleets. It was asked if the relationships between AMO cycles were related to local-scale environmental conditions (e.g. temperature and salinity) within other parts of the West BFT range. The authors stated that this work is ongoing and that they are examining micro-scale environmental patterning and then scaling up their analysis to broader spatial and temporal scales.

4.3 Specify sensitivity runs

East (VPA)

No further sensitivity runs were requested.

West (VPA)
The Group reviewed a sensitivity run for West BFT that included the 2018 Canadian Gulf of St. Lawrence acoustic index data point for VPA (SCRS/P/2020/027) and Stock Synthesis (SCRS/P/2020/023) provided during the meeting.

It was highlighted that the inclusion of that index's 2018 value would result in reduced recruitment estimates for the last 10 years (**Figure 16**) as 2010 showed a less increasing trend while the estimated recruitments were lower than those in the 2017 assessment (Anon., 2017b).

West (Stock Synthesis)

The Group noted that removal of the 2018 Gulf of St. Lawrence acoustic data point from the base case had no conspicuous impact on Stock Synthesis model outputs (**Figure 17**). This was not the case in the West BFT VPA model (**Figure 16**), which was influenced by the removal of the 2018 acoustic data point. The Group noted that the Stock Synthesis model includes a broader range of indices that are not included in the VPA, including two indices that show increases in abundance for West BFT in 2018. It was further noted that the two models have underlying structural differences and consistent sensitivities should not be expected.

4.4 Specify projection setting

East (VPA)

Discussion on projection specifications involved reviewing which specifications were used in 2017 and which should be used for the 2020 update. In the 2017 assessment, a 6-year period (2006-2011) was used to calculate a geometric mean as a patch value to replace the recruitment values in the 4 terminal years of the model (then 2012-2015) (Anon., 2017c, BFTE-4. Outlook). During the 2017 assessment (Anon., 2017b), the 6-year average was chosen as it aligned with medium recruitment scenario in the 2017 assessment.

The Group discussed the problem of the highly variable recruitment estimates in the post-2007 retrospectives, which also have very different perceptions on scale and pattern between Run0 and Run135 (**Figure 5**). In Run135, which is accepted for use in the 2020 update, the retrospective pattern shows highly variable recruitment estimates in scale and pattern (**Figure 5**). The change in perception of the scale of recruitment post-2007, in Run135 vs Run0, will result in a much higher geometric mean if the 6-year average used in 2017 was updated. The Group felt that simply updating the 6-year average would not be appropriate for the 2020 patch value update (**Figure 18**) as these differences in post-2007 recruitment estimates between the 2017 assessment (Anon., 2017b) and Run135 were not robust to small changes in the recent input data. The Group agreed to use the 1968-2007 period to calculate the geometric mean level of recruitment, which approximated a medium recruitment scenario, as a more robust approach in these circumstances. Due to a lack of confidence in the recent estimates of recruitment, the Group also agreed to use this "patch" value for all years after 2007 (2008-2018), as opposed to the last 4 years as used in the 2017 assessment. Further specific details on the methodology employed for the patch will be provided in documents produced by the assessment team.

The bullets below provide the values used in 2017 and what was agreed for use in the 2020 update.

- Choose which statistic is used to calculate the patch value
  - In 2017: geometric mean
  - Agreed to use in 2020: geometric mean
- Choose which period is used to calculate this statistic
  - In 2017: assessment statistic calculated over 2006-2011 period
  - Agreed to use in 2020: 1968-2007
- Choose the years that will have their recruitment values replaced with the patch value
  - In 2017: 4 years replaced (2012-2015)
  - Agreed to replace in 2020: 11-year patch (2008-2018)
- Choose years for selectivity
  - In 2017: 2012-2014
  - Agreed to use in 2020: 2015-2017
- Catches for 2019 and 2020
  - Based on Rec (2019 = 32,240 t and 2020 = 36,000 t)
    - The Secretariat has preliminary catch information which indicates that CPCs caught the 32,240 TAC set for 2019, so it appears this will be a good estimate of realized catches
    - 2020 needs to be thought through due to the possible impacts of COVID
- Iterations: 500 bootstraps (80% CI)

- Catch intervals to be used in projections: 1,000 t intervals (in Kobe matrices and chicken feet) between 18,000-50,000 t annual catch and F0.1 value (in chicken feet)
- Projection years: 2021-2025

West (VPA)

The Group agreed on the following projections to be run by the next meeting in July:

- Future recruitment
- Continuity "rule" would use 2010 to 2015 (i.e. 6-yr mean excluding terminal 3- yrs)
- 3-year recruitment patch, 2016 to 2018
- Selectivity for projection years: geomean of 2015 to 2017
- Quota scenarios range 0 to 3,500 by 250 mt increments, plus constant exploitation at $F_{0.1}$ and $F_{current}$ (exactly as defined in 2017 assessment report (Anon., 2017b))
- Provisional 2019 and 2020* catches at 2,350 mt
- Project constant catch scenarios, F0.1, F_CURRENT for 2021-2025 for the assessment report and 2021-2023 for the executive summary
- Many of these specifications will also apply to SS projections. If not specifically noted in Section 4.4, ensure that the analysts work together to provide comparable projection advice between the VPA and SS3 models
- 80% confidence intervals reported on fishery status and projected yields
- K2SM based on 1,000 combined model bootstraps
    - 500 bootstraps from the VPA
    - 250 from SS young spawn scenario
    - 250 from SS older spawn scenario

The Group agreed that the projections will follow the same pattern as in 2017: three-year projections will be shown for management advice and additional projections extended to five years will be reviewed in July. The Group requested that the analytical team provide a full assessment report to be reviewed during the July meeting. Exactly the same figures for the base model diagnostics and results as in 2017 should be produced for consistency and comparison to the 2017 assessment (Anon., 2017b).

The request of the Commission to modify the TAC intervals was noted by the participants. However, and mainly due to the lack of time to appropriately deal with this issue, it was postponed to the July meeting, in which the projections will still be shown using the current TAC intervals. If any decision can be taken during the July meeting, the assessors will work intersessionally to have the outputs with the new TAC intervals by September.

West (Stock Synthesis)

The Group agreed to use the tentative settings for preliminary projection found in **Table 4** to generate Kobe phase plots and Kobe II projection matrices. Consistent with the West VPA, future recruitment will be set based on a 6-year geometric mean (2010-2015) and F-current will be set using a 3-year geometric mean (2016-2018). Several options exist for evaluating uncertainty. Parametric bootstrapping was used in 2017, whereas since then, a multivariate lognormal approximation (MVLN) method (Winker *et al*., 2019) has been developed and applied to both YFT and BET. The Group recommended that the MVLN method should be compared to the bootstrapping method used in the last assessment. The Group discussed whether projection tables would continue to use 250 t increments. It was decided that 250 t intervals would be used in the short term and decisions regarding finer scale intervals (e.g. 100 t) will be made during the July or September 2020 BFTSG meeting.

5. Summary of developments on ABTMSE

A presentation was provided by the Group Co-Chair, giving an overall perspective of the current situation. As a consequence of the uncertainties arising from the pandemic situation, the MSE process could experience a delay relative to what was planned in the roadmap agreed in 2019. Everybody is now learning to work by correspondence, and it remains to be seen how fast the work can progress in these circumstances. At this stage, the MSE process continues its development to the extent possible and the

situation should become clearer in the second half of the year. For the time being, the BFTSG has an interim OM (Operating Model) grid to use as a framework and is starting the process of evaluating projection behaviour and associated CMP (Candidate Management Procedure) performance using the updated MSE software package.

The presentation noted that the BFT-MSETG meeting that took place in February 2020 (Anon., 2020b, *in press*) provided major breakthroughs in relation to the OMs, as it managed to successfully address important difficulties found in earlier meetings (in particular, the problem of not being able to determine the abundance scale of the BFT populations satisfactorily in the conditioning process, which had been identified in the BFT-MSETG meeting of July 2019 (Anon., 2020a , *in press*) and caused delays in the process). The OM interim grid arising from the February 2020 BFT-MSETG meeting (Anon., 2020b, *in press*) has five uncertainty axes: two of them ("Recruitment"; "Mortality/Maturity") are as before, but three are either new ("Likelihood weight" of the length composition data; "Scale": average SSB in the West and East areas) or modified from the previous specification ("Mixing": average proportion of western stock biomass in the East area). These three axes bracket the mixing scenarios previously discussed by the BFTSG provide a range of scale values for the East and West areas around those found in previous stock assessments conducted by the BFTSG, and offer an alternative with improved fit to the length composition data. Details of the work and conclusions of the February BFT-MSETG meeting can be found in its report (Anon., 2020b, *in press*).

The resulting OMs (96 in all for the interim reference grid) no longer suffer from the difficulties encountered before, and consequently the complicated likelihood weighting assignment that was being explored during the Autumn of 2019 is no longer needed. The OMs are not perfect representations of the BFT populations and their spatial dynamics, but the aim in MSE is not to find a best OM/representation of these dynamics, but rather to have a set of OMs that span the range of realities considered plausible, so as to provide an appropriate testing framework for CMPs. The aim is to have a framework that allows the development of a CMP to provide catch limit recommendations that should work well in practice in achieving management objectives.

In the CMP development and simulation work that will occur over the next few months, further understanding of OM behaviour in projections will be gained. For example, it may be found that certain uncertainty axes of the interim OM grid do not have an impact on the performance-based ranking of CMPs; it may then be decided to drop such axes from the OM grid. Conversely, some cases currently classified as robustness tests may be found to have a clear impact on CMP performance and, hence, may end up being elevated to the final OM grid.

One of the main questions raised by the Group was whether it is acceptable to examine CMP behaviour to help finalize the OM grid, or whether the OM grid should be finalized before any CMP behaviour is examined. MSE experts responded that being very strict in the MSE process in trying to finalize the OM grid exclusively based on the perceived plausibility of OMs normally led to extremely lengthy discussions among the scientists without clear conclusions, and, hence, that a more practical and effective MSE process would consider reviewing some aspects of the interim OM grid based on whether or not they were found in simulations to matter for CMP performance. This discussion was postponed until later in the meeting (see Section 7.1, later in this report).

5.1 Summary of new package

The BFT MSE contractor provided a presentation (SCRS/P/2020/025) and noted that the outcome of the February BFT-MSETG meeting (Anon., 2020b, *in press*) had been very positive. The Group Co-Chair then presented the report of the February BFT-MSETG meeting in some detail. A summary of both presentations and the discussion that followed in the meeting is provided below.

Interim OM grid:
The interim OM grid spans a range of stock status, absolute stock size and recent stock trajectories. The grid covers wide ranges and provides a challenging and diverse set of OMs on which to test the CMPs.

Robustness tests:
Robustness tests and their priority order as proposed by the February BFT-MSETG meeting (Anon., 2020b, *in press*) were explained, noting that three tests were assigned the highest priority: senescence; the western stock growth curve applied to the eastern stock; and Brazilian catches allocated to the East area. The

February BFT-MSETG meeting considered that some of the previous robustness tests are no longer needed, because the issues they intended to address have been dealt with or no longer arise for the new interim OM grid. The Group was invited to make proposals for robustness tests, but none were received.

In relation to the questions raised in the Group discussion about why certain robustness tests were no longer considered necessary or the desirability of undertaking additional robustness tests to those already entertained, MSE experts explained that the interim OM grid is already very large (96 OMs) and that there is already a very long list of robustness tests. In the next stage of the MSE process, some changes to the interim OM grid and robustness tests (and their priorities) will be considered, but it has been impossible to code all these tests in the limited time available to the contractor.

It was also explained that the interim OM grid and the robustness tests have been built based on extremes, which are assumed to encompass the range of plausible realities. Depending on the sensitivity or lack thereof found in CMP performance, it may be that at a later stage in the MSE process a more detailed examination of some of the uncertainty axes, or of certain robustness tests, could occur (e.g., by including additional intermediate cases between the extremes). However, it would be premature and overly complicated to do this now.

Indices for potential use in CMPs:
The work of the February BFT-MSETG meeting (Anon., 2020b, *in press*) concerning indices for potential use in CMPs and how they will be simulated in the MSE software was presented. Full details are available in Section 5.1 of the February BFT-MSETG meeting report.

The Canadian acoustic survey in the Gulf of Saint Lawrence is simulated in the MSE software, but index developers have advised that it should not be used in CMPs because of the uncertainties surrounding the index in recent years (and, hence, in how it should be projected into the future). Therefore, future simulated values of this index should not be used in CMPs.

The aim is to generate future index data with similar properties (for the residuals of the observed log(index) values relative to the OM-fitted values) to those observed in the past. Tables at the end of the February BFT-MSETG meeting report contain four statistical measures examined during that meeting. It was decided to generate future index values with the same SD (standard deviation) and AC (autocorrelation) values as observed in the past (for the residuals of the log(index)).

The February BFT-MSETG meeting eliminated only one index based on the observed properties of the historical residuals. This was the Canadian RR CPUE index for the Gulf of Saint Lawrence, which showed very poor behaviour across all four statistics examined and, consequently, it was considered very difficult to generate future simulated data replicating its past properties (in particular, biases). The February BFT-MSETG (Anon., 2020b, *in press*) meeting proposed simulating all other indices in the MSE software for potential use in CMPs. This decision was questioned by the Group, in particular in relation to the French aerial survey whose historical residuals have a very high SD (around 0.8). MSE experts noted that, all other things being equal, indices with higher SD values will contribute less useful information to CMPs than indices with lower SD values. This, however, should appear as an emergent feature of the CMP performance, so that there is no need to exclude such indices a priori. It was also noted that all indices in the East area correspond to short series, which, consequently, does not allow many choices of indices for the East area.

Updated MSE software package (version 6.6.12):
The package now contains the 96 OMs in the interim reference grid. It simulates indices for potential use in CMPs, and future recruitment values, according to the specifications of the February BFT-MSETG meeting (Anon., 2020b, *in press*). New functions have been incorporated in the package to facilitate visual understanding of projection performance and to help critically review the outputs (including the correctness of the code and the realism of the outputs). These include further checks of the historical years (to ensure that the MSE package exactly replicates the OM fits, which were conducted separately in ADMB) and displays of simulated indices as well as recruitment and SSB development during future projection years. Additional developments are planned for the near future, including the addition of the main robustness OMs and catch projections by fleet (as previously requested by some BFTSG members).

It was noted that in the interpretation of MSE results, careful attention is needed to ensure the meaning of the term "distribution" is clear in different output displays. In MSE, this term arises in several contexts, with two main ones being: a distribution from a single OM (based on the output from multiple stochastic simulations from a same OM); a distribution across OMs (e.g. when a single value summarizes some output from an OM, and the distribution of such values across the different OMs is considered).

A question arose concerning Recruitment level 3 (a factor on one of the uncertainty axes in the interim grid of OMs), which is specified as a change in recruitment regime occurring "10 years into the projection period". Because the projection period in the MSE software starts in 2017, this means that this shift occurs in 2026 in the software, and there was some concern that this might no longer be appropriate given that any potential MP will only be used to set the TAC for 2022 (at best) or 2023. MSE experts noted that, for appropriate testing of CMPs, the regime shift could essentially occur at any time, because the aim is to test whether the CMPs are able to react to it appropriately whenever it occurs; the main constraint from a practical point of view is that the regime shift should not occur too late into the projection period, because then there may not be sufficient years left in the projection to appropriately evaluate resource reaction to a CMP after the regime shift has occurred. At a later stage in the process, it might be appropriate to test final CMPs for their behaviour for different choices for the future year when the shift takes place. It was also noted that the MSE is expected to be reviewed every 5 years, at which point the OMs should be reconditioned including the latest data. It was agreed that the actual years (e.g. 2026 in this case) should be more clearly specified in the MSE description and outputs.

5.2 Update of trial specification document (TSD)

The main TSD (Trial Specification Document) changes made for this meeting refer to the specifications for simulation of indices and future recruitment **(Appendix 5)**. On page 1 of the TSD it is noted that this is a work in progress. Writing the TSD accurately (mathematically) for an MSE model of the complexity of that developed for BFT is an enormous exercise and it has not been possible to achieve this in full so far; further work is needed. The following was agreed:

- Members were invited to review the TSD and to send comments by e-mail to the BFT MSE contractor.
- The TSD section on performance measures needs to be developed further, providing additional detail and a simple but clear explanation of each of the performance measures. This is necessary for the correct interpretation of MSE outputs.

It was also noted that performance measures proposed by the 2019 Intersessional meeting of Panel 2 (Anon., 2019a, Appendix 8) should be included in the TSD. However, the list of measures previously indicated by Panel 2 is very large. Once trade-offs start to emerge from the CMP testing process, it will be possible to discuss them with Panel 2. In the experience of MSE experts, once managers start to understand the trade-offs, it becomes much easier for them to identify the key performance measures, and this should help focus and finalize the MSE work.

The BFTSG will also have to find ways to condense the extremely extensive and complex set of outputs from the MSE into a format that can be useful and understandable for the managers that need to make choices concerning a potential BFT MP.

5.3 MSE process for approval of various elements by the BFTSG

The final OM reference set (or "grid") needs to be approved and adopted by the BFTSG, and the same is the case for the plausibility weighting process to be used for the OMs (see Section 7 of the February BFT-MSETG meeting report (Anon., 2020b, *in press*) for options to address this issue). These processes may require face-to-face meetings (given the subtleties of some of the aspects that need to be discussed, particularly for the OM plausibility weighting process, and potential language complications when working remotely). The Group Co-Chair advised that the rules of engagement and timing of these processes have not yet been decided, in view of the uncertainties surrounding the current pandemic situation. The SCRS Chair added that, in the current circumstances, it is necessary to remain flexible and to try to adapt to the evolving situation as it unfolds.

6. Assessment initial results, if available

6.1 East (VPA)

Based on the diagnostics (see Section 4.1), the Group agreed to use Run135 as base case for the 2020 update assessment. It corrected the severe retrospective bias in SSB as seen in the 2017 model and relatively improved fits to several of the indices over the 2017 model. Weaknesses of this run include that, like the 2017 model, recent recruitments show substantial uncertainty, as evidenced by the lack of robustness to small changes in the input data. This suggest that all recruitment estimates from 2008 onwards are unreliable for projections. Hence the Group decided to replace 11 years of recruitments (2008-2018) with the long-term geometric mean recruitment (1968-2007), rather than relying upon recent estimates (see Section 4.4). The implications of this "patch" are that the model has very little information on recent status and trend; this may make projections undesirable for providing advice. These projections will be conducted intersessionally and reviewed in July.

Research recommendations included the evaluation of larval/juvenile survival indices to assess strong and weak cohort signals, particularly during the period 2008 to the current year. In addition, several of the treatments of the purse seine size composition need to be revisited, notably the "other PS" CAA that is a combination of multiple fleets and may reflect changing in sampling and not changes in the fishery and the PS Croatia size samples which, before the stereo-cameras were implemented, were based on harvest size and not size at caging.

6.2 West (VPA and Stock Synthesis)

The Group agreed that a 2020 base case model of VPA and Stock Synthesis is the strict update of the 2017 stock assessment (Anon., 2017b), however the 2018 Canadian Gulf of St. Lawrence acoustic index data point is omitted (see Section 2). A sensitivity run that included the 2018 data point was also provided (see Section 4.3).

Both western models (VPA and Stock Synthesis) exhibit similar historical trends and magnitude to the 2017 models, respectively (see Section 4.2). In the West, there are conflicting indices. The VPA model shows some preference for a higher F-ratio compared to the fixed values from 2017. For Stock Synthesis the lack of a strong stock recruitment relationship and a certain lack of fit to the composition data were identified as other diagnostic weaknesses. For both models, these diagnostic patterns were identified but not deemed to be of a magnitude that would require straying from a strict update assessment. Hence these weaknesses, as well as a number of issues identified in the 2017 assessment (Anon., 2017b) still need to be addressed in benchmark assessments. Nonetheless, the lack of severe bias in retrospective estimates of recruitment, in sharp contrast to the East VPA and the similarity between VPA and Stock Synthesis, indicates that the recent recruitments are consistently estimated, which was not the case with the East BFT VPA. Further, the congruence in SSB scale and recruitment for VPA and Stock Synthesis across modelling platforms with substantially different structure and basic data inputs gives some further confidence in their reliability.

Some members expressed their wish to check the S-R plots at some point during the meeting and these will be included in papers presented for the July meeting that document model results and preliminary projections. While 2017 and current assessment advice does not use inferences from a stock-recruitment relationship for benchmark setting, it is useful to view the inclusion of an additional 3 years of data.

7. Initial experiences with the new ABFT MSE PACKAGE 6.6.12

7.1 Presentations from CMP developers

SCRS/2020/075 was presented. Key issues explored in the document were: 1. whether the interim OM grid can indicate which uncertainty axes matter and which do not for CMP performance; 2. whether using a CMP, including feedback control, provides improvement relative to the poor performance displayed by a simple "Current Catch" (continue with the current catch into the future) policy. CMPs for the East and West areas were developed using a weighted average of several indices and a constant harvest ratio (HR) approach. Five CMPs were considered: two without feedback control ("Zero Catch", "Current Catch") and three with feedback control (corresponding to 3 different values for the HR in East and West areas: "1-1", "0.75-0.75", "0.5-0.5").

In relation to the first question, the document examined the distribution of the performance statistic Br30 (SSB in projection year 30 divided by dynamic SSB_{MSY}, per stock) across the 96 OMs under a "Current Catch" policy. Separately for each stock, the 96 OMs were ordered according to the resulting Br30 values and the 30 worst OMs identified, aiming to gain an understanding of how the different levels of the uncertainty axes in the OM grid impacted on performance. The main feature that emerged was that Recruitment level 2 (a single historic and future recruitment regime for each stock) was associated with the worst Br30 results for both stocks, very often resulting in stock extinction. The conclusion drawn was that a CMP with feedback control was required to avoid this undesirable outcome.

In relation to the second question, distributions (histograms) of Br30 and avC30 (average catch over the 30 projection years, per area) over the 96 OMs showed an improvement in Br30 values (i.e. higher Br30 values and generally avoiding stock extinction) when going from a "Current Catch" policy to a CMP with feedback control. Comparing different CMPs, the expected trade-off between avC30 and Br30 emerged: CMPs with higher HR (harvest ratio) resulted in higher avC30 but in lower Br30. Another relevant aspect was that when examining the distribution of Br30 across the 96 OMs, the CMPs with feedback control displayed a narrower distribution (lower SD) than the "Current Catch" policy; this is a desirable robustness property gained as a consequence of the feedback control. The next step (not yet developed) would be tuning, i.e. adjusting for a preferred catch-depletion trade-off.

Discussion:

The Group discussed the issues highlighted in the presentation. The main concerns raised related to whether this kind of analysis provides any information about the appropriateness of OMs, particularly about their plausibility, and whether looking at such results at this stage could lead to bias in the selection of the OM grid, as one begins to glean the impact of different OMs on estimated future catch and stock status.

The Group converged towards an understanding that this process does not inform about the plausibility of OMs, but may serve to identify the aspects (uncertainty axes) that are most influential on CMP performance and those that have no effect. This can facilitate the process of discarding some non-influential uncertainty axes and their associated OMs, because they will not be useful for distinguishing (ranking) among the different CMPs, and hence focusing subsequent attention on the more influential aspects.

In summary, there are two different issues here: OM plausibility and which aspects matter (and do not matter) for CMP performance, and the view was expressed (by some) that the efficiency of the CMP selection process makes it necessary to consider both aspects interactively. The Group Co-Chair concluded that the key balance that needs to be achieved is between adherence to strict principles (decide upon OM grid and weightings before seeing results) and the practical necessity of viewing some performance metrics to evaluate what matters, and noted that it is important to develop a structured process to facilitate future discussion and decision-making.

It was agreed that a worked-through example would be provided to facilitate further discussion of this issue, and this is included in SCRS/2020/079).

SCRS/P/2020/024 presented empirical and model-based CMPs that are being developed by the authors of this presentation. Since the OMs in the interim grid are meant to represent the range of plausible "real-worlds", their aim was to find a CMP that works acceptably across the entire OM grid. A cluster analysis was performed on the trajectories of the conditioned OMs and a number (k) of clusters were identified, as well as the OM corresponding to the medoid[1] of each cluster.

A CMP was considered, where the TAC was a weighted-average of k (number of clusters) TACs; each of the k individual TACs is designed to be responsive to the biomass trends of the OMs in a particular cluster, and the weights could be proportional to the number of OMs in each cluster. In the example presented, each of the k individual TACs was obtained by fitting a delay-difference model using the stock biomass of the corresponding cluster-medoid OM and the actual simulated indices as data, and subsequently applying a Harvest Control Rule (HCR), e.g. a sliding rule based on F_{MSY} as estimated from the delay-difference models and the estimated biomass (as well as potential caps on the total catch allowed), to the outputs of that delay-

[1] The "medoid" is a mathematically representative object (e.g. a biomass time series) within a set of such objects; it has the least average dissimilarity to all other objects in the set (www.wolframalpha.com).

difference model. Some initial results from several alternative CMPs of this type were presented, focusing on stock depletion and average catch statistics over some future periods of years. For each of these alternative CMPs, Br30 and C10 values over all 96 OMs in the interim grid were displayed in increasing order; the rank order of CMPs generally stays the same for all OMs. A way to plot a performance measure (e.g. Br30) across the 96 OMs in the grid and for multiple CMPs was presented, aiming to provide a tool that could help uncover patterns regarding uncertainty axes in the OM grid that may or may not impact on CMP performance. However, the complexity of the output makes it very challenging to see patterns and draw conclusions.

Discussion:

The approach was generally found to be interesting and further development was encouraged.

The Group discussed how the plots shown could be used to try and understand which aspects of the OMs had little impact on CMP performance. The plots were very useful to gain a first impression of patterns, but in order to make further progress on reducing the number of OMs, it could be useful to apply a method such as a response surface analysis or fitting a GLM to the outputs.

The issue of 'omniscience' in constructing CMPs was raised. It was agreed to postpone discussion of this matter until the July meeting where the discussion would focus on defining the problem, so that the BFTSG could reach a conclusion on that point (also taking into account information that CMP developers could bring).

The need to be careful with terminology was stressed. In particular, the BFTSG agreed to generally use the term "refining CMPs" for improvements made to CMPs, and to reserve the term "tuning CMPs" for the very specific aspect of adjusting CMPs to achieve a particular value in the catch-depletion trade-off space.

SCRS/P/2020/022 was a quick exercise carried out by some EU scientists to test the new MSE package and their CMP developed last year. This is an index-based CMP that uses the median of 4 indices (each of which has been normalised to a mean value of 1 over the years 2015-2018) for each of the East and West areas. Initial "targets" were taken as 0.75 for East and 1 for West; when the median of the 4 indices in recent years is above the target, the TAC increases, whereas the opposite occurs if the median in recent years is below the target. Some exploration was conducted drawing on the ideas in SCRS/2020/075. In particular, the Br30 values for the 96 OMs in the interim grid were plotted for each of the East and West areas in increasing order. In each case, the OMs for which Br30<1 were extracted and the proportion in which the different factors in each uncertainty axis appeared was examined. The results suggest that "Scale", "Recruitment level", and possibly "Mixing" are influential axes. Some more detailed results were shown for some OMs. Everything at this stage corresponds to preliminary choices only, and considerable further development is to be expected.

Several clarification questions were asked, and some suggestions were offered regarding more effective ways of using indicators in CMPs, in particular related to the degree of responsiveness of different indicators, as well as the impact of their auto correlation.

Several questions were raised concerning the fact of not using the GBYP aerial survey index. The argument for that was that authors tried to find a balance between the indices used in terms of numbers of indices, number of indices related to juvenile and adult fractions of the population, number of indices related to the pure Mediterranean and the Gulf of Mexico areas and the mixed Atlantic area. A concern was highlighted about the possibility of losing this year for some of the indices due to the exceptional circumstances this year due to the pandemic (so a previously unexpected gap may occur in some indices), and the most appropriate way to deal with this situation (e.g. by trying to design CMPs that are robust to this). MSE experts felt that this usually makes little difference to CMP performance and that it is best discussed later in the MSE process, as part of Exceptional Circumstances.

SCRS/2020/079 (first part of document) presents an exercise, illustrating an approach to assess which uncertainty axes in the interim grid of OMs matter (and which do not) for CMP performance. The approach is applied to the "Current Catch" policy, which has no feedback control, and two CMPs with feedback control, namely the "0.75-0.75" and "0.5-0.5" HR (harvest ratio) CMPs from SCRS/2020/075, which correspond to higher and lower (in relatively terms) exploitation intensities, respectively. The aim of the document is to present a generic approach and ideas, not to focus on the specific detail of these CMPs.

The main results are provided in Table 1 of the document, which repeats the same exercise already presented in SCRS/2020/075 for the performance statistics Br30 and AvC30, with the difference that the values now presented in each row of Table 1 were calculated not from the entire set of 96 OMs in the interim grid, but from a subset corresponding to the level indicated in the first column of the table for the relevant uncertainty axis. For example, the first row of the table presents results for the subset of OMs that use "L" (i.e. low) in the uncertainty axis "likelihood weight for the length composition data" (i.e. 48 OMs), whereas the second row of the table does the same for the OMs that use "H" (i.e. high) in that axis (i.e. the other 48 OMs). The next four rows of the table repeat the exercise for the levels of the "Scale" axis, and so on. The right-most column of the "EAST" and "WEST" blocks of Table 1 is a new calculation and is the maximum difference between the "Br30 Median" values in the table within each uncertainty axis. The values in this column are the key ones used to examine which axes matter for CMP performance: the larger the value for an uncertainty axis the more significant the impact of this axis on CMP performance. The values in that column are presented graphically in Figure 1 of the document, to help visually identify patterns. The left-most panels of that figure correspond to the "Current Catch" policy (no feedback control), and the other panels to results from the CMPs with feedback control.

Figure 1 of the document indicates that one of the uncertainty axes (the one plotted by the right-most bar in each panel of the figure) generally has higher values than the other uncertainty axes and, thus, is a relevant axis that matters for CMP performance. It should also be noted that the values obtained for this axis decrease when going from the "Current Catch" policy to a CMP with feedback control (as expected, because feedback control should be able to react, hopefully in an appropriate manner, to the varying situations encountered in future years, hence counteracting them in ways that result in the increased stability of the resource; this is the robustness property expected under feedback control). Focusing on the CMPs with feedback control, the figure shows that two of the uncertainty axes consistently display very low values and hence scarcely matter for CMP performance, whereas the other three uncertainty axes definitely do matter. The presenter also noted that caution should be exercised when examining this type of figure, as occasionally it could give the wrong impression, particularly for badly designed CMPs (such as the "Current Catch" policy, which has no feedback control, and for which the figure suggests that the uncertainty axis corresponding to the left-most bar does not matter in the West, whereas this axis is seen to matter when more appropriate CMPs, with feedback control, are examined).

A further result confirming the important role played by feedback control in CMPs is the fact that, in Table 2 of the document, the range of (Br30 median) values resulting from CMPs with feedback control was generally small and the extent of the ranges less than the difference between having or not having feedback control.

Two main conclusions emerging from this analysis were (text taken directly from the document):

- Decisions about which uncertainty axes "matter" in terms of CMP performance can be made before any decision on desired CMP tuning with respect to final abundance targets need be reached.

- The approach above can provide a basis for assessing which uncertainty axes "matter", and hence also a basis for perhaps deleting some of these in the current interim grid and also considering replacing them with others. However, before this can be done reliably, the CMPs must be refined further by their developers to improve the robustness of their performances to closer to that which might be possible for their eventual final forms.

Discussion:

The approach outlined in the presentation was found to be very useful.

One point clarified was that if an uncertainty axis was determined to not be influential on CMP performance and ended up being eliminated from the OM grid, this would actually imply that a single level for that axis would be used in all OMs. How exactly that single level would be chosen is something to be discussed later in the MSE process; at this point, the main task is to identify uncertainty axes that could be targets for elimination.

A main concern raised about the results presented was that the primary "Mixing" western-stock-mixing-into-the-eastern-area axis did not have an impact (Figure 1 of the document presented) on CMP performance, which could be interpreted as counterintuitive based upon previous considerations of the importance of mixing. Certain members of the Group responded that some surprises always emerged in this type of process and that robustness tests could be conducted to explore alternative scenarios (in this case, alternative mixing hypotheses) for their potential impact on CMP performance; depending on the results of these robustness tests, some uncertainty axes not currently represented in the OM grid could later be included. Nevertheless, concern remained among BFTSG members about the fact that the current OM interim grid might not be capturing the uncertainties associated with mixing sufficiently. The Group Co-Chair commented that, whereas this is obviously a concern, it is necessary that CMP developers experiment with the current interim grid so as to have a firmer basis on which to draw conclusions. Clearly, results that are unexpected *a priori*, particularly on mixing, will have to be explained (both for the scientists, and for managers and stakeholders); for example, it could be that the feedback control built into CMPs is having the effect of counterbalancing the variation which alternative mixing scenarios would otherwise create. In such situations, the Group would be well-served by having explanations for these or, in such situations that defy simple explanation, confidence that such behaviour does not matter for the end results, a process outlined in Section 7.4.

Another concern raised about the approach presented is that it focuses exclusively on "main effects" and does not examine "interactions" between uncertainty axes, and that, moreover, it may be too subjective. Another methodology (e.g. GLM analysis) would be useful to explore the results in more detail and in a more objective manner; this type of analysis has been used on large grids of OMs in some Indian Ocean MSE work. The presenter responded that such additional analyses could be useful and would be welcome, but that one should also be mindful not to overcomplicate things; the simple approach provided in the presentation already pointed towards some clear messages and, whereas sufficient analysis should be conducted in order to make the necessary decisions for the MSE, care should also be exercised not to use extra resources unless necessary.

Another comment made about the approach was that alternative calculations to those shown in Table 1 and Figure 1 of the document presented could have been useful. For example, concerning the first two rows of Table 1, instead of separately looking at the "L" and the "H" OMs, the analysis could have been based on paired OMs, where a pair of OMs would have equal levels of all uncertainty axes except for the "likelihood weight" uncertainty axis, which would be "L" for one element of the pair and "H" for the other one. Another suggestion made was to look at Br30 differences not in terms of absolute values but relative to the overall magnitude of the Br30 values (within each uncertainty axis). The presenter noted that these were interesting ideas and worth considering at later stages, although he expected that at the broad level of consideration they were likely to provide similar results to those from the approach shown in the presentation; he stressed that, whereas there are obviously multiple ways in which the extensive amount of output could be examined, the key is to find a balance between the resources spent on examining the output in alternative ways and the actual level of analysis required to make decisions for the MSE to progress.

Finally, it was noted that the analysis shown in the presentation focused exclusively on Br30 and that, whereas this performance measure is important, Commissioners have consistently identified other objectives (relative to safety, yield and stability) as also being very relevant. The presenter noted that the one trade-off that strongly dominates in CMP performance results is the catch-recovery trade-off, i.e. the catch taken over a period of time in relation to the abundance of the resource at the end of that period of time. It was agreed to continue this discussion later in the meeting where it is addressed in Section 7.5.

7.2 Further future package extensions: Shiny apps for plotting and summarizing, etc

In addition to new expected features in the MSE software package, the MSE contractor indicated that the next round of documentation will have information about how to run multiple MSEs in parallel, and there will be functions for plotting the output from multiple MSEs against each other, as well as facilities for the standard reporting of MSE results. The package will also include the main robustness OMs and catch projections by fleet (as previously requested by some BFTSG members).

7.3 Plans for further interactions amongst developers

Substantial work by CMP developers and interaction between them will be needed in the next few months (certainly between now and September, and continuing after that date). A substantial part of that interaction could occur via remote meetings.

7.4 Discussions regarding the path forward for the MSE process

The second part of SCRS/2020/079 proposes a process to follow in order to move the MSE process forward, given an interim OM grid and very preliminary CMP results. The proposal provided a valuable template for development of the BFTSG path forward.

Discussion on SCRS/2020/079:

The Group Co-Chair noted that, whereas the timelines will likely be hard to accomplish given the circumstances, the presentation was very helpful in terms of proposing a path forward for the tasks needed to conclude the MSE process. He also noted that the BFTSG members are now learning to make progress in remote meetings and this should help MSE work to progress (e.g. interaction between CMP developers) even in the current circumstances.

Many meeting participants agreed that the timeline seemed optimistic but found the description (and proposal) of tasks very useful. The meeting then focused on discussing the actual steps proposed, rather than the timeline.

Several participants noted that the seven-step process described (SCRS/2020/079) in the presentation was different (and more complicated) from what they had understood last year and their expectation coming into this meeting.

The "Mixing" issue noted earlier with regards to the current interim OM grid was raised again in the context of whether the current grid is sufficiently representative of the major uncertainties to be able to start the seven-step process on this basis. As noted earlier, it was agreed that further exploration of the interim OM grid with CMPs was necessary and that unexpected results, particularly with regards to mixing, would need to be properly understood (for example, examining the projected catches of western fish in East area fisheries) and explained to stakeholders. Again, it was noted that appropriate robustness tests with alternative mixing scenarios will be conducted and their results evaluated, in particular, examining if they have an impact on CMP performance.

A main issue causing concern for several meeting participants was that the focus now seems to have strongly shifted from evaluating OMs based on their perceived plausibility to evaluating them in terms of their impact on CMP performance. It was noted that, with 96 OMs currently in the interim grid, it is practically impossible to review OM conditioning results on an individual OM basis (the procedure followed in the past), but that the February BFT-MSETG meeting (Anon., 2020b, *in press*) considered the results of OM conditioning and almost every reality check had been passed; however, some participants commented that they had spent time reviewing individual OM reports and that some problems remained.

After discussion, the BFTSG converged towards the view that, although the plausibility of OMs is obviously important and the BFTSG should continue to be cognizant of potential issues, focusing at this stage on what matters for CMP performance will provide valuable insights for subsequent work (including the possibility of further reviewing the OM conditioning outputs for a possibly reduced interim grid) and to help focus the plausibility weighting assignment work (to be conducted in Step 6 of the proposed process) on the most influential aspects.

It was also agreed that a subgroup of meeting participants that have spent time reviewing the individual OM reports, would review the spreadsheet of previously identified issues to see if they were still relevant and propose a robustness test, or at least highlight the issue identified, to the meeting before Day 6. Meeting participants were also reminded that they need to consider the robustness tests and priority order indicated in the February BFT-MSETG meeting (Anon., 2020b, *in press*) report and the additional ones in the TSD, and if they wish to raise any issues they should do so before Day 6 (**Table 5**, Current and additional robustness tests).

In response to these discussions the WG developed the following proposal for a **Path forward for the BFT MSE process**.

An initial version of the 'path forward' (Section 7.5, below) was reviewed by the Group with a number of concerns raised regarding the proposed path and the MSE process in general. The Group raised several questions regarding the path forward for the MSE process.

First among them was a comment that the precise timing of events after the July meeting need not be very specific, a view held by the Group, given current uncertainties.

A general concern with the process of decision making was raised that related to the request (Anon., 2019d) by the SCRS that the contractor provide: "A time/feasibility estimate for OMs with reduced dimensionality" and "If it cannot be provided by February, and, if the Reference OM set is not adopted by the April 2020 meeting, the BFTSG should reconsider the costs and benefits of reducing the complexity of the current MSE framework."

During the SCRS meeting, the BFT MSE contractor noted by email that it: "would require over 1 year of additional work, fundamental redevelopment of the model, requiring iteration with the full BFT-MSETG to re-specify how indices relate to new strata and how electronic tag data information is to be reinterpreted in the fitting process with no guarantee of a better outcome." This time estimate made it unfeasible for the contractor to conduct this restructuring by this current meeting, particularly given the priority placed upon reconditioning the current OMs. However, the full message related to the complexity involved in simplification may not have been effectively conveyed by the Group Co-Chair. Another omission was that the decision to reconsider the costs and benefits of reducing the complexity of the OMs was not presented to the BFTSG.

Simplification of the OMs is a topic that has received substantial attention recently. Many in the BFTSG share a concern that the models may be too complex to diagnose. Hence this concern needed to be discussed and the path outlined in the 2019 SCRS report (Anon., 2019c) required that a decision on this matter be taken by the BFTSG at this meeting. The Group Co-Chair asked the Group if it wanted to proceed with the current 7-area operating model structure or to move towards a reduced structure.

While the Group did not have the time to consider the full costs-benefits of a reduction in complexity, substantive breakthroughs in previously identified issues with the OMs and their passing of nearly all red-faced tests was noted as a key development. Additionally, it was not clear that simplified OMs would be necessarily easier to construct or that they would have the necessary structure for evaluation of complex management questions. In the end, The Group decided to continue with the current OM structure and its associated robustness tests.

An additional concern was raised that this meeting did not provide a thorough review of the output of the OMs from the February BFT-MSETG meeting (Anon., 2020b, *in press*) and that it did not adequately address intersessional feedback. These concerns were noted and, in response, the proposal for a structured, curated venue for addressing issues at the July BFTSG meeting is outlined in part 2 of Section 7.5, below.

The Group also raised concerns regarding how the process may ask too much of participants, both of the BFTSG and the BFT MSE contractor. One concern noted that the process outlined in part 2 of Section 7.5 places a high burden on the BFTSG to demonstrate problems with the OMs. Conversely, it was noted that asking the BFT MSE contractor to evaluate every possible issue is time and cost prohibitive. A solution may be the above-mentioned curated process at the July BFTSG meeting, which will occur after participants and CMP developers have had several weeks to work with the OMs.

The Atlantic Bluefin ecosystem of multiple stocks with complex population dynamics and numerous fisheries involving 53 ICCAT contracting parties may necessitate unique approaches to foster inclusivity, communication and the smooth functioning of the process.

7.5 Path forward for the BFT MSE process

This path forward represents an aspirational proposal to conclude the MSE work in time for the adoption of an MP in autumn 2021. Due to the current pandemic that has precluded the in-person meetings originally deemed necessary to complete tasks such as adoption and plausibility weighting of the reference grid, the MSE process will probably experience a delay until these and other critical tasks can either be completed through in-person meetings or arrangements are made to conduct decision making through remote meetings. Utilization of remote meetings may well assist in the completion of many of these tasks. Regardless of the logistical details, the following seven steps remain necessary to complete the MSE process.

1) [MAY-JUNE] Developers work intersessionally
Objective: To determine which of the interim grid axes "matter" most in terms of the relative extent to which they impact key performance statistics (e.g. Br30, …)

- Developers refine their CMPs by using the present interim grid together with the robustness trials deemed most important
- Developers submit refined CMPs to contractor
- Suite of CMPs (need not be final) used to screen OMs to determine what 'matters' according to the criteria outlined in Section 7.5, below.
- Contractor runs suite of CMPs and updates the shiny app developed earlier to be able to provide tabulated output for July meeting

2) [JULY] July BFTSG intersessional meeting
Objective: To examine combinations of OMs and CMPs so as to identify behaviour that might not match earlier perceptions or that could be difficult to explain. To be conducted during the first 1-2 days of the July BFTSG meeting

- BFTSG participants work intersessionally to identify problematic behaviour
- Participants work with the contractor and BFT-MSETG to conduct due diligence by reviewing combinations of OMs and CMPs to provide illustrations that support assertions of such problematic behaviour (e.g. plots/tables or package outputs)
- Where the issue raised relates to facilitating understanding, participants consult with BFT-MSETG members (who include CMP developers) to summarise explanations appropriately. Where further model modifications may be necessary, this same process is used to formulate such instances in the form of OM hypotheses
- BFTSG then judiciously considers whether it is worth investing the time and expense of the following process:
    - The proposing participants work with the BFT MSE contractor to specify the model (e.g. which parameter values can be adjusted in conditioning to achieve the desired change in behaviour)
    - The OM will be evaluated first to determine whether it deteriorates the quality of the OM conditioning sufficiently to render the alternative unacceptable.
    - If this first test does not create diagnostic problems, the new OMs are run under several CMPs to see whether they are influential on the results under the proposed method outlined in (1) above and proposed in SCRS/2020/079.
    - If influential, then such OMs could be considered for elevation to a high priority robustness test or even to the interim grid.

3) [JUNE] Developers virtual meeting
Objectives: Developers (through the BFT-MSETG) consider and propose a small set (probably three) of interim "recovery" targets to which to tune their refined CMPs for the purpose of convenient and comparable presentation of initial results. This is the process of "development tuning" (see Section 10 and Appendix 8 of BFT-MSETG meeting report from April 2018 (Anon., 2019b)).

a. These would need to span the range likely to be of interest to the Commission.
b. They might (for ease of implementation) most readily be defined in terms of the median value in a stochastic implementation of one "centrally performing" OM in the interim grid of a biomass performance statistic such as Br30.

4) [JULY] BFTSG virtual meeting

Objectives:

1. Review of steps taken by developers and BFT-MSETG (Anon., 2020b, *in press*) in steps 1) -3) above.
2. Coordination of CMP developers work for summarization of part (1) results to facilitate BFTSG consideration of axes that matter, and mutual interaction to help them further refine their CMPs.
3. Agreement on possible modification of the axes to be included in the interim grid, and designation of the most important robustness trials to be considered in further work.
4. CMP developers will review individual CMP characteristics and their ability to achieve interim objectives and propose a best alternative/hybrid CMP.

5) [TBD] Developers work intersessionally

a) Developers refine their CMPs further, now using the development tunings in 3), and also taking into account a fuller range of performance statistics.

b) Based on these results, developers (through the BFT-MSETG) also suggest a smaller set of performance statistics which capture the key differences in performance (independent of the main catch-recovery trade-off reflected by the different development tuning choices) amongst their CMPs. (Note that in practice, many performance statistics are so highly positively correlated that their further consideration is not helpful in distinguishing qualitative differences in CMP performance.)

6) [TBD - possibly some issues below could be addressed in the September BFTSG meeting]
Considering the developers' results from the process above, proceed to discuss the following topics and to agree on the combination of possibilities it wishes to proceed in some iterative manner, which will in turn be linked to further runs of the CMPs (note that in this process, prior consultation of BFTSG members with their Commissioners/managers/stakeholders will be helpful to inform deliberations):

c. Agreement to possibly modified values for interim development tuning targets.

d. Agreement to a possibly modified "smaller set" of performance statistics (note that in many past actual cases, the ultimate selection of an MP has involved consideration of performance for at most only two performance statistics).

e. Agreement on the final reference grid of OMs – note that the constituent uncertainty axes are unlikely to change at this time, but the values of the factors on these axes might change from the initial "extremes" to somewhat more central options. What decisions are made at this stage will depend also on decisions on the matters raised in the bullet below.

f. Agreement on a system for assignment of plausibilities/plausibility weightings to different OMs, and on how to utilise these in developing recommendations to stakeholder groups (see Section 7 of the report of the February 2020 BFT-MSETG meeting) (Anon., 2020b, *in press*).

g. Cull of the CMPs surviving to this stage of the process to a small number (maybe two or at most three) for which to present results to stakeholders and refine further on the basis of their feedback. Note that this final set might include tested "CMP-combinations", which involve taking the TAC outputs from, say, two different CMPs and implementing a catch which is some weighted average of the two.

h. Agreement on the range of tunings for which to present such results (these could be identical to the development tunings – they are NOT final advice on a final tuning range for which to present results would come from iterative interaction with stakeholders and ultimately be provided by the Commission).

7) [TBD] BFTSG meeting

At this stage of the process, the BFTSG develops a presentation of results of a first set of CMP options through the SCRS to Panel 2/ the Commission for their response, and will also develop a proposed specification of an iterative interactive process of dialogue with stakeholders to lead to a final proposal of options for an MP to be made to the Commission.

7.6 A general methodology for evaluating what operating model axes "matter"

a) A general approach (as in SCRS/2020/079) — screen a series of CMPs across the full suite 96 stochastic OMs, calculate the maximum percent difference in the medians for key performance metrics Br30 and AvC30. In the case of screening of robustness tests, a similar approach will be used for a specified set of OMs; this may require that the sets specified in the February 2020 BFT-MSETG meeting (Anon., 2020b, *in press*) are somewhat enlarged (a process on which the contractor will be advised by the developers).

b) Appropriate performance metrics from the full grid of OMs must be provided by several refined CMPs to provide robustness to the analysis and to examine the interplay between CMPs and the axes of uncertainty.

c) Summarization, tabular, visual (as in SCRS/2020/079) and using the Shiny app which is to be developed further by contractor

d) Additional performance statistics to Br30 — when providing their refined CMPs to the consultant, the developers should each also have investigated alternative performance statistics and provide suggestions to the consultant about which (if any) should also be included in his/her summary report.

e) Employ a statistical (GLM) methodology to evaluate the significance of grid axes and their factor levels. A statistical approach will help to determine whether there are significant interactions that render interpretation of the main effects incomplete or misleading. This approach can include consideration of the effect of CMPs and their potential interaction with the axes of uncertainty. The response to which the model is fit can be Br30, AvC30 or some other performance metric. This approach has the advantage of an established model selection and validation process as well as helpful methods for effect visualization.

f) Definition of 'matters' Axes that matter survive the model selection process and thereafter account for a greater than appropriate percentage of the variability in the response (Br30, AvC30, etc.). This determination must include consideration for the effect of any significant interactions that can make the assessment of main effects incomplete. Careful consideration of the full implications of dropping or condensing an axis would most effectively be done at the July BFTSG meeting where results from the screening process are available.

g) Axes that do not pass the 'matter' criteria could be condensed requiring that a default (central) input must be selected to be fixed across all surviving OMs.

h) Plausibility and "influence": The process to determine which OM axes are influential is to be followed for plausibility weighting at a later date after the process has been agreed upon.

8. Initiation of working group on growth in farms

The Group Co-Chair explained to the rest of the members that the purpose of the creation of the Sub-Group on growth in farms (BFT GF-SG) is that the SCRS response to the Commission (Rec. 19-04, para 28) will be more comprehensive with the participation and advice of all the scientist involved in the related ongoing studies.

The Group accepted the creation of the specific Sub-Group on growth in farms after the approval of the Terms of Reference. Furthermore, it was stated that this Sub-Group will inform the Group about all work, and that the final decision-making authority will remain the Group.

8.1 Identify team lead

Dr Deguara volunteered to lead and coordinate the growth in farms Sub-Group. The Group recognized and accepted the proposal.

8.2 Draft terms of reference and coordination plan

Dr Deguara summarized to the Group the Terms of Reference.

Terms of Reference
"In response to the request by the Commission as per para 28 of ICCAT Rec. 19-04, the mandate of the Sub-Group on growth in farms (BFT GF-SG) is to present information to the Bluefin tuna Species Group to enable an updated Growth Table to be formulated by consensus with the aim of providing as comprehensive an answer as possible to the Commission."

Furthermore, he presented the Working Plan agreed in a preliminary online meeting of this Sub-Group on growth in farms, organised on the 11 May 2020 and established in its report.

General questions raised by the Group were clarified. A question was asked regarding the relationship of this Sub-Group to GBYP activities. The GBYP is specifically tasked with carrying out or overseeing several experiments related to growth in farms and some data analyses and is a key member of the Sub-Group and will contribute essential data. But it is not alone in this task as, importantly, many individual CPCs also have information or ongoing experiments that will provide valuable data for this response. Ultimately, as the response to the commission is a task and responsibility of the SCRS, this BFT GF-SG has been created to assist the Group to provide the SCRS with a consensus response.

Regarding several technical questions particularly related to how the studies were going to proceed or whether individual growth could be affected by tagging, it was highlighted that there is no definitive answer yet and these are key issues to be considered in the activities planned by this Sub-Group. As is the case with any Bluefin Tuna Species Group Sub-Group, participation is open to the all members of the Group. Interested parties are requested to contact the Sub-Group Chair.

9. General discussion of GBYP matters

The Group Co-Chair provided an overview presentation on some of the central themes of the programme that require particular attention under 4 broad categories: Aerial surveys, tagging, workshops and other projects. COVID-19 has imposed a number of restrictions on programme activities and has had a major impact on GBYP projects and workshops. Many aspects of the programme have had to be postponed to some point in the future, while others have been cancelled for the current year or Phase.

9.1 Aerial survey

All GBYP aerial surveys in the Mediterranean Sea have been cancelled in 2020 due to COVID-19 restrictions, due to the timing (May/June) and international logistics. In addition, a recent report (Cañadas and Vázquez, 2020) on the reanalysis of the 2010-2019 Aerial Survey data identified several areas of uncertainty that directly affect how the index was standardized.

Over the time series, several changes have occurred to improve sampling methodology and design. Base data sets have been also recently refined. Consequently, data analysis has incorporated those changes as well as new methods of standardization. These changes challenged the development of indices over the time series and resulted in several interpretations on how the data should be standardized. Assuming that the new and revised process is the best interpretation, the perception of stock abundance from this index is very different to the one used in the MSE process.

The primary objective of the overview was to make the BFTSG aware of these results and consider recommendations on how to move forward. A series of questions were provided to guide the decision-making process, including the possibility of contracting an external advisor. Three decision options were proposed for September: 1. Continue as designed with modifications/enhancements to address the uncertainties, 2. Cancel portions of survey (maintain some regions), and 3. Cancel the GBYP aerial surveys.

Discussion:

The discussion centred around the recent re-analysis of the GBYP Med aerial surveys and the changes in the standardized index, compared with previous standardizations, and what to do about the inconsistencies. The Group Co-Chair pointed out that a decision on aerial surveys is not required until September, but wanted feedback on options for the interim. The reason behind the notable change in the standardization results must be addressed in detail. Given the significant change in the index, the BFTSG was invited to go through the latest report provided by the contractor with an overview of the analysis, results and

conclusions of the most recent re-analysis.
One of the issues which arose during discussion was "Can the Group resolve the concerns internally or do they need to recruit an external advisor/reviewer?". It was noted that the GBYP Steering Committee (SC) had tried to capture some of these concerns but that they had not been resolved. If the latter approach is chosen, the challenge will be finding an external expert familiar with tuna aerial surveys as most are associated with marine mammal surveys. Furthermore, there is a lot of information in the existing report and it may be better to use internal experts for the review. However, given the importance of the survey and the time required to conduct a proper review it may be advantageous to hire an external reviewer. A decision will have to be made in the relatively near future. The cost of an external review may not be significant compared to the cost of the survey, especially considering the workload of those who should be involved. There is also a requirement to develop, circulate and discuss the terms of reference for the external review, if this will be the choice.

Decision options for consideration in September include cancelling, or continuing with the GBYP aerial survey in full (e.g. all Mediterranean) or in part (e.g. specific areas), and considering further improvements to deal with the index development and standardization methods. Participants were reminded that the Mediterranean is a very important BFT spawning area with few indicators of abundance. Consequently, it is imperative to explore all tools available to revise or enhance this index before discarding the aerial survey. It was noted that the available W-Med larval survey is limited in coverage area to the Balearic Sea. Other possibilities include looking at an aggregated index for the entire Mediterranean (which is currently available in the new analysis) or evaluating if the index is worth the information it is providing using MSE. The index may be so variable that it does not contribute much to inform management, possibly due to differential abundance in each area in each year. Other questions to be addressed include the most adequate treatment and statistics, if the data are representative, if the school size affects standardization (i.e., introduces a bias), and if the problems can be resolved without starting a new time series. Solutions will come at a cost which also needs to be considered. Several suggestions were made for starting new time series that involved surveying with drones and the use of satellite images to detect schools and school size, but it was reported that these proposals have been already examined, discussed and rejected in previous years.

Conclusions and suggestions

The Group recommends examining the aerial survey standardization in detail.

Continue the internal review through July and September with a decision required in September on how to proceed.

Develop Terms of Reference for the external review, if selected as the option to address the issue.

Translate additional documentation on the GBYP review report and circulate (post) to BFTSG members as soon as possible.

9.2 Tagging

There is a need to develop strategies for the very short-term (summer/autumn 2020) future in the absence of the planned workshop to develop more long-term strategies. Opportunistic tagging was postponed in the Canary Islands and tagging in the eastern Mediterranean was postponed until 2021 to allow for tagging of migratory individuals. Other tagging opportunities are possibly available through agreements with other ongoing CPCs e-tagging programs

9.3 Workshops (tagging, close-kin, larval)

All workshops were postponed due to COVID-19. The close-kin workshop was to resume consideration of the feasibility of the application of this method. The Group preferred that the GBYP workshops be held in late 2020 if possible, giving priority to the close-kin workshop due to logistics and planning issues. The Group expressed that these workshops could be handled through webinars.

9.4 Progress on other GBYP projects

Other GBYP projects continue as planned: Data recovery, Modelling, Biological studies. Up-to-date information on each of these projects is available at https://www.iccat.int/gbyp/en/biostu.asp

10. Workplan leading to the September assessment

10.1 Assessment and Executive Summary

1. Preliminary projections (using 2019 and 2020 TAC) will be distributed to the WG (6th July)
2. Assessment advice and elaboration of an Executive Summary draft (20-28th July)
3. As usual evaluate 2019 realized TAC but not project with values.
4. Final advice for 2021 and 2022 (maybe 2023 TAC (for 21 September))
5. Specifications:
    i. West: July Assessment report projections to 5 years, executive summary projections 3 years
    ii. East: July Assessment report projections to 5 years, executive summary projections 3 years
6. What-if projection scenarios of implications of 2020 catches that might be under the TACs due to COVID-19-related reductions in effort (to be conducted by analysts)
7. Updates of indicators through 2019 for Executive summary (final 2019 indices due in September)

10.2 Responses to the Commission

a) Growth in farms, see Sub-Group on growth in farms (BFT GF-SG) terms of reference, Sub-Group to meet intersessionally by webinar

b) Response regarding 'update catch rates by fleet segment and gear'.

10.3 MSE

See BFT MSE Path forward (Section 7.5)

11. Other matters

Given time constraints, scientific papers not specifically related to the stock assessment or MSE were considered in other matters. SCRS/2020/017 was noted by the Group Co-Chair, but the Group did not have time to formally review it and it can be on the agenda for the July meeting. SCRS/2020/057 was presented by its author.

11.1 SCRS/2020/017, Nottestad et al.

The Group received a brief mention of SCRS/2020/017, which describes fishing capacity and catch rates for Norwegian purse seine vessels from 2014 to 2019. The paper was presented to Panel 2 in March and should be considered in greater detail by the BFTSG at the July meeting, where the response to the Commission on fishing capacity will be considered in greater detail.

11.2 SCRS/2020/057, Di_Natale

This paper represents an annotated bibliography of the vast volume of literature (now more than 2040 papers) from Italian researchers on the subject of bluefin tuna dating from antiquity to the present time. This work opens up much of the previously difficult-to-find literature on the species and will be of value to the BFTSG and researchers of bluefin tuna. The Group thanked the author for this contribution.

12. Adoption of the report

The Report of the first 2020 ICCAT intersessional meeting of the Bluefin tuna Species Group was adopted. Drs Gordoa and Walter thanked the participants and the Secretariat for their hard work and collaboration to finalise the assessment and the report on time. The meeting was adjourned.

References

Anon. 2017a. Report of the 2017 ICCAT bluefin tuna data preparatory meeting (Madrid, Spain, 6-11 March 2017). ICCAT Collect. Vol. Sci. Pap. 74(6): 2268-2371.

Anon. 2017b. Report of the 2017 ICCAT bluefin stock assessment meeting (Madrid, Spain, 20-28 July 2017). ICCAT Collect. Vol. Sci. Pap. 74(6): 2372-2535.

Anon. 2017c. Report for biennial period, 2016-17, Part II – Vol. 2. 427 pp.

Anon. 2019a. Report of the 2019 Intersessional meeting of Panel 2 (Madrid, Spain, 4-7 March 2020). 137pp. https://www.iccat.int/Documents/Meetings/Docs/2019/REPORTS/2019_PA2_ENG.pdf

Anon. 2019b. Report of the 2018 ICCAT bluefin tuna Species Group MSE intersessional meeting (Madrid, Spain, 16-20 April 2018). ICCAT Collect. Vol. Sci. Pap. 75(6): 1056-1159.

Anon. 2019c. Report for Biennial Period, 2018-19, Part II, Vol. 2. 470 pp.

Anon. 2019d. Atlantic bluefin tuna Species Group meeting summary report (Madrid, Spain, 25-28 September 2018). ICCAT Collect. Vol. Sci. Pap. 76(2): 686-692.

Anon. 2020a. *In press*. Report of the 2019 second intersessional meeting of the ICCAT bluefin tuna MSE Technical Group (St. Andrews, Canada, 23-27 July 2019). ICCAT Collect. Vol. Sci. Pap. 76(1): 165-201.

Anon. 2020b. *In press*. Report of the 2020 intersessional meeting of the ICCAT bluefin tuna MSE Technical Group meeting (Madrid, Spain, 24-28 February 2020). SCRS/2020/001.

Ailloud, L.E., Lauretta, M.V., Hoenig, J.M., Hanke, A.R., Golet, W.J., Allman, R., and Siskey, M.R. 2017. Improving growth estimates for western Atlantic Bluefin tuna using an integrated modelling approach. Fish. Res. 191: 17-24.

Cañadas A., and Vázquez J.A. 2020. Final report of 2010-2019 Aerial Survey Data Re-Analysis under the Atlantic-wide Research Programme for Bluefin tuna (Phase 9) Short-term contract ICCAT GBYP 07/2019

Cort, J.L. 1991. Age and growth of the Bluefin Tuna, *Thunnus thynnus* (L.) of the Northwest Atlantic. ICCAT. SCRS/1990/66. Collect. Vol. Sci. Pap. ICCAT, 35 (2): 213-230.

Winker, H., J. Walter, M., Cardinale, D. Fu. 2019. A multivariate lognormal Monte-Carlo approach for estimating structural uncertainty about the stock status and future projections for Indian Ocean Yellowfin tuna. IOTC-2019-WPM10-17.