Stock Assessment: Independent peer review of the Pacific halibut stock assessment

PREPARED BY: IPHC SECRETARIAT (D. WILSON, 23 OCTOBER 2019)

PURPOSE

To provide the Commission with an opportunity to further consider the independent peer review report of the IPHC Stock Assessment for Pacific halibut.

BACKGROUND

The Commission directed the IPHC Secretariat via Commission decisions **AM095-Rec.10** and **IPHC-2019-ID001** (shown below) to:

> **95th Session of the IPHC Annual Meeting (AM095) – 1 February 2019**
> **AM095–Rec.10 (para. 129)** "*The Commission **RECOMMENDED** that the IPHC Secretariat develop terms of reference for a consultant to undertake a peer review of the IPHC Pacific halibut stock assessment, for implementation in early 2019. The terms of reference and budget shall be endorsed by the Commission inter-sessionally.*"
>
> **2019 Inter-sessional decision – 17 April 2019**
> **IPHC-2019-ID001**: The Commission **ENDORSED** the "*Open call for expressions of interest: Independent peer reviewer for the IPHC stock assessment*"

The report by the independent consultant was provided to the Commission on 2 August 2019, via IPHC Circular 2019-16.

DISCUSSION

The report by the independent peer reviewer, Dr Kevin Stokes, is provided at **Appendix I**, and is also available on the Stock Assessment page of the IPHC website under the 'Peer Review' tab for transparency and accountability purposes: https://www.iphc.int/management/science-and-research/stock-assessment. A direct link to the pdf is also provided below:

https://www.iphc.int/uploads/pdf/sa/2019/stokes_2019-independent_peer_review_for_the_2019_iphc_stock_assessment.pdf

The review will be considered at the Commission's upcoming Work Meeting (18-19 September 2019), and also by the IPHC's Scientific Review Board at its 15th Session from the 24-26 September 2019.

RECOMMENDATION/S

That the Commission **NOTE** paper IPHC-2019-IM095-08 which provided the Commission with an opportunity to further consider the independent peer review of the IPHC Stock Assessment for Pacific halibut.

APPENDICES

Appendix A: Independent peer review of the Pacific halibut stock assessment (K. Stokes)

APPENDIX I

Independent Peer Review for the 2019 IPHC Stock Assessment

Prepared by Kevin Stokes
August 2019

Summary

This report reviews the in-development 2019 full stock assessment of Pacific Halibut being conducted by the Secretariat of the International Pacific Halibut Commission (IPHC). The stock assessment is updated annually and undergoes full assessment every 5 years. The last full assessment was in 2014. The basis for the full stock assessment should be completed by September 2019 for final review by the IPHC Scientific Review Board (SPB) before its application to all updated data in December 2019 and provision of science-based risk assessments to the IPHC for decision-making in early 2020.

This review covers the full spectrum of stock assessment related matters and is guided by the terms of reference set out by the IPHC. The potential scope is large and the review attempts to focus on key matters, based on the terms of reference and discussion with the IPHC Secretariat. The review included a site visit to the IPHC in Seattle which overlapped with an SRB meeting. The SRB has separately provided feedback to the Secretariat on the in-development stock assessment.

Pacific halibut has been exploited for over a century along the North American west coast by IPHC members (USA and Canada). Commercial fisheries started in the 19th century along the west coast but even before 1920 had expanded to the Gulf of Alaska. The majority of the stock is distributed in Alaskan waters and over time the commercial fisheries in Alaska have come to dominate mortalities. Since the 1960s, bycatch in commercial Alaskan trawl fisheries has grown. Pafiic halibut provides important subsistence catches and has also been increasingly taken by guided and non-guided recreational fisheries since the late 1970s. Despite the wide array of fishery sectors, data on mortalities and catch rates are generally of a high quality to inform stock assessment. Some minor areas of concern are noted in this review, including the section on research priorities.

Biological data from commercial fisheries are generally sound though as fish are landed dressed, sampling at ports is critical. A key issue is determination of commercial catch sex ratios. Work by the IPHC to determine sex ratios using port sampling and genetic analyses is in hand and new data have already been considered in the in-development stock assessment. This work is important and may need to continue beyond the initial 2 year program.

The IPHC operates a dedicated and extensive annual setline survey which provides the stock assessment with critical information on Pacific halibut abundance and distribution as well as with biological data. Exploratory work to improve the survey has been in progress since 2014 and should come to fruition in late 2019/early 2020 to inform the 2020 design. The survey, which uses a large number of member country commercial vessels annually, is outstanding by any measure and provides not just critical inputs to the stock assessment but also an important platform for ongoing and agile research to understand Pacific halibut biology and ecology. State of the art approaches are used to analyse survey data and provide high quality indices and

other data to the stock assessment. The survey is critical in that provides information on fish that will enter the fishery three or four years later.

The stock assessment is conducted using the Stock Synthesis framework and is carried out by world class analysts, supported within the IPHC by statistics and biology teams and by the independent SRB, and embedded in the fertile Seattle stock assessment and methods community. The quality of analysis if excellent and aimed purposefully at providing science-based risk assessment to support IPHC decision-making.

Individual stock assessment models have been developed iteratively over many years but have settled since the last full assessment to include four structurally different models that are fitted in a two-way cross to Long (i.e., full history) or Short (i.e., since 1992) data series and to Coastwide (i.e., as a single area) or AAF (i.e., Areas-as-Fleets). The models use different approaches to fixing or estimating natural mortality, selectivity, and environmental factors. The rationales provided for the model development are credible and robust based on historical analyses, data availability, and utility. All models are individually fit using state of the art manual, iterative tuning techniques which are well explained. As an in-development assessment, final tuning will be required once the assessment approach is agreed and final 2019 data become available. The in-development assessment considers addition or replacement of models for the final assessment. This review finds the four models a good basis for providing a consistent, robust and credible risk assessment to the IPHC in early 2020. Especially given the progress being made on Management Strategy Evaluation by the IPHC Secretariat, for possible implementation of agreed mortality-setting rules by 2021, major changes to the existing set of stock assessment models is not encouraged.

The provision of risk assessment advice to the IPHC uses all four, structurally different models, in a way which is slightly unconventional. Most stock assessment-based advice is based on a single assessment and associated sensitivity runs to portray uncertainty. While that approach may provide risk assessments that include uncertainty associated with data and model fitting to data, it does not address uncertainty due to the structural differences between models - all of which are valid. Selecting a single model as a basis for risk assessment puts a key part of the risk decision in to the science process rather than the IPHC Annual Meeting process. In order to separate risk decisions in science and policy to the greatest extent possible, the IPHC approach is to assess risks associated with any decisions on future mortalities using an ensemble of all four models. Selection of the four models is rational and science-based and use of all four removes the necessity to focus on any one model.

Of course, different models could be selected and risk assessments could be affected. The rationales for model development are, however, science based and credible. In order to provide a consistent basis for advice this review concludes that continued use of the four individual models is appropriate. This leaves open the issue of whether the four models might be weighted equally, as in recent years, or differentially. There is no right way to weight the models and even equal weighting is arbitrary. Equal weighting also makes models with lower biomass scales

influential in assessing risks. The issue of weighting is considered in the review and at this stage it is advised to maintain equal weighting.

The IPHC is conducting Management Strategy Evaluation which is likely to result in adoption of rules for setting mortalities in 20121. Once implemented, it is possible the need for annual stock assessment updates will be removed. This would provide time to analysts to explore more fully a range of important issues such as automated tuning of individual models, alternative individual models to account for structural uncertainty, weighting of models within the ensemble, use of Bayesian approaches (also impacting on ensemble weighting options). All of these are considered in the review as well as all other research priorities outlined in IPHC stock assessment and data update papers.

Background: ToR, Process, and relationship to IPHC Performance Review

Terms of Reference (ToR) for this stock assessment (SA) review are intentionally wide, providing scope for discussion and focus as deemed appropriate on the *stock assessment process, methods and reporting*. Nevertheless, *specific topics that should be addressed* fall in the following categories:

1) Aspects of data collection and analysis.
2) Aspects of individual model development. [Aspects of developing individual models to consider for including in the ensemble.]
3) The collection of models contributing to the ensemble, and the methods for combining/weighting the results.
4) Comments on research priorities or avenues for data, model or management advice development as appropriate.
5) Comments on the document and background material provided for the review.

The review is also required to *clearly delineate between tactical changes to be considered for the current (2019) stock assessment and research avenues for future work.*

The review was carried out remotely but benefited from an informal site visit from 17-20 June 2019 to meet IPHC staff, discuss a range of SA issues, identify key SA documents, and understand the IPHC website structure and content. The site visit also provided an opportunity to discuss science processes, to be reported on separately as input to the 2nd Performance Review (PR) of the IPHC (PRIPHC02). The site visit was not initially planned and I am grateful to the IPHC staff who made time and contributed to it.

The IPHC SA is undertaken within the Secretariat by dedicated science staff. The primary focus of this review is the SA *per se*, conducted by the *Quantitative Sciences Branch*. Inputs to the SA and aspects of research planning and prioritisation, however, also require consideration of work carried out by the *Biological & Ecosystem Sciences Branch* and the *Fisheries Statistics &*

Services Branch. During the site visit, four presentations were provided by the three IPHC Branches as background and to aid discussion. The presentations used were the same as given to the 1st session of PRIPHC02; they are available online at: https://www.iphc.int/venues/details/2nd-performance-review-of-the-iphc-priphc02-1st-session.

The last full SA of Pacific halibut was in 2015 with updates in 2016, 2017 and 2018. The in-development SA now being reviewed (the 2019 assessment) is the first weigh point in the first full assessment since 2015. Expectations about the SA are provided in the report of the 13th Session of the Scientific Review Board (SRB; IPHC, 2018): *A full assessment analysis and review is planned for 2019, which will allow more in-depth investigation and model-based evaluation of the new and/or revised data. Progress continues on the reevaluation of whale depredation accounting in the Fishery Independent Setline Survey time-series, as well as the sex-ratio of the commercial catch in 2017; both products are anticipated in February 2019. That analysis will also allow for an in-depth exploration of data weighting, parameterization of time-varying processes and other modelling approaches implemented in the four Pacific halibut models comprising the stock assessment ensemble.*

The key SA document for the review is Stewart and Hicks (2019). As a first weigh point in the 2019 process, the paper describes and reports on preliminary analyses conducted during the development of the 2019 SA. It includes consideration of new data; bridging from the previous assessment, including consideration of issues noted by the SRB; initial individual model weighting; and initial ensemble modelling. While it superficially provides indications for status in 2019, these should be treated cautiously given the imminent addition of full 2019 survey, fishery and other data, and potentially any changes in models used.

The IPHC SA process includes two SRB meetings annually; the preliminary SA report is presented and considered in June each year and feedback from the SRB is used in development of the final SA that is presented to the SRB in mid-late September. Completed current year data are then used in final model runs and development of decision tables to be used by the Commission. This review is timed to allow any findings to be considered alongside comments made by the SRB in the report of its 14th session. Stewart and Hicks (2019) has in fact already been considered by the 14th Session of the SRB which met from 24-26 June 2019 (IPHC, 2019a). The SRB made just three requests of the SA team: one regarding the IPHC setline survey and two regarding the SA modelling. These are commented on below.

ToR bullet 5 (*Comments on the document and background material provided for the review*) can be dealt with quickly and simply at the outset. The SA paper by Stewart and Hicks (2019) is notable for its careful and logical elaboration of the in-development SA. It is unusually and exceptionally clear with a focus on explaining why as well as how models have been developed - from an historical perspective, given data, and in the IPHC decision-making context. While many SA documents focus on model fitting, Stewart and Hicks (2019) is about modelling but with full consideration of model fitting nested appropriately, comprehensively and clearly. It is an excellent document but for review needs to be read in conjunction with Stewart and Webster

(2019) which elaborates on data available for the SA. It also needs to be considered in the context of its purpose which is to provide a scientifically rigorous, but value-free, risk assessment to aid the Commission in its annual deliberations.

In addition to the in-development SA document, a wide range of papers and materials were made available for the review in electronic form, either in advance, during the informal site visit, or through the IPHC website. In advance, these included detailed input and output files for the individual models (see ToR bullet 2) used in the ensemble (see ToR bullet 3); the excellent, annually updated, overview of data sources up to November 2018 (Stewart and Webster, 2019; ToR bullet 1); previous model documentation; and relevant papers/manuscripts on the assessment, most notably as relevant to ToR bullets 2 and 3. The overall quality of documentation from all IPHC sources is of the highest quality with exceptional care taken in preparation.

Data Collection and Analysis

ToR bullet 1: Aspects of data collection and analysis.

Stewart and Webster (2019) provides an annual update of data as of November 2018. The paper is clear and comprehensive in scope as of November 2018, identifying data changes and additions but not repeating methods as outlined in previous documents. Data as relevant to the SA development, including bridging and weighting, have also been summarised in Stewart and Hicks (2019). During the site visit for the SA review, a number of relevant presentations were made (as also made to the PRIPHC02, see above).

Full review of all data sources is beyond the scope of this review. Review, for example, of fisheries statistics collection or the Fisheries Independent Setline Survey (FISS) could be standalone. Only key aspects of data collection and analysis are commented upon here. Stewart and Webster (2019) note a number of data sources for potential future analyses and relevant research projects. All of these are also included in a wider list of research priorities outlined by Stewart and Hicks (2019). These are all commented on in the section below on *Research priorities, Biological understanding* or *Research priorities, Data related research*.

The data available for Pacific halibut SA are unusual in that they span a long period of time and comprise both high quality fishery dependent and independent sources which are well documented and understood. The fishery dependent and independent sources are remarkably coherent. For example, the comparison between the FISS over-32” WPUE and commercial WPUE from 1995 onwards can be seen clearly in slides 10 and 11 of IPHC (2019b) and between FISS indices and commercial WPUE reported in Stewart and Webster (2019). While the sex ratios of the FISS and commercial catch are different, the trends and scales are neverthless suggestive of a high degree of consistency between the indices, reflected also in the good fits to all indices in the individual models reported in Stewart and Hicks (2019). Comparisons of compositional data from different sources also appear consistent. Of course,

the SA needs to balance compositional and other data with indices and to fit complex selectivities, estimate mortality, etc, but the coherence overall gives reassurance that the final SA should be able to provide i) a robust view of the Pacific halibut stock status, and ii) a sound basis for risk assessment related to future mortalities. It is usual in SA to need to make hard decisions about data weighting in individual models which go beyond rigorous statistical considerations. With such coherent data there is a reasonable *a priori* expectation that weighting choices might be less important than is often the case. Also, with such coherent data it is reasonable *a priori* to expect between-models correlation of trends and estimates of variance on status metrics and forecasts (see below on ensemble modelling).

Pacific halibut is caught by an array of sectors across a wide geographic range and in two national jurisdictions. Even with the majority of the catches being taken in directed setline fisheries, fisheries data collection and preparation is therefore complex.The IPHC has its own observers but relies necessarily on its member states' national data collection programs for fisheries-dependent data that feed into the SA. In discussion with IPHC staff, this seemed to be regarded as a weakness, but it is normal for cross-boundary stocks managed by RFMOs and the overall quality of mortality data does seem to be good. The IPHC clearly works directly with fisheries and has good relationships that enhance data collection and understanding of issues. IPHC staff visit ports and vessels and the annual use of multiple commercial fishers for the FISS is a means not just to collect high quality data but also to develop relationships that underpin confidence in wider data collection. Ongoing access at ports, e.g for fin clipping to determine sex ratios in commercial catches, is a good example. Confidence in following regulations and reporting is also created in, e.g., USA complete lack of head-off landings in 2017 and 2018 following regulatory change in early 2017 (IPHC, 2017 para 48).

IPHC (2019b) and Stewart and Webster (2019) provide a summary of the multiple fishery components by sector and area. My overall impression is that while the data collection systems could always be better specifically for halibut, they of course are designed for multiple species with a wide range of constraints. Given those constraints, there seems in the main documentation to be general satisfaction that the nature and extent of mortality is reasonably captured. The lack of sensitivity testing in historic and current SA suggests it is not regarded as a major uncertainty. However, some concerns are implied at *Research priorities, Data related research* items 10 and 11 which propose (10) reanalysis of historical bycatch mortalities and age frequencies, and (11) investigation of variances and errors in the scale of mortality estimates; these concerns are commented on below. IPHC (2019c) notes a number of concerns related to recreational, subsistence and bycatch fisheries. Considering concerns expressed by both IPHC (2019c) and Stewart and Webster (2019), only one common issue seems to emerge - the low level of observer coverage in directed fisheries in Alaska, with none for vessels less than 40', leading to inaccurate fish weights and age-distributions for discarded fish. The Alaska commercial fishery mortality is a large percentage of the total (circa 50%) and of the Alaska fishery the discard percentage is of the order of 5%. While 5% of 50% may seem small, information on fish below the MLS is important in determining selectivities and providing information on recruitment to the SA. It is beyond the scope of this review to recommend

improving observer coverage by a member state but this is clearly one aspect of mortality estimation where improved information would be useful and could improve credibility of the SA.

One potential unaccounted mortality component is whale depradation in the commercial fisheries, as has been observed, quantified and explored for the FISS (see below). This is not mentioned in Stewart and Hicks (2019), even under *Research priorities*, or other documents but was raised in discussion during the site visit. The possible scale and nature is unclear, as is whether it might (or not) be important in the risk assessments provided for decision-making. While discarding could create an unaccounted mortality of smaller fish that might impact estimated future risks, depradation by whales of the same scale as discarding might be important to estimated status and/or future risks depending on its nature (i.e., size of fish taken or trends). Generally, for all stock assessments, consistent biases in unaccounted mortalities should "come out in the wash" if fishing practices remain consistent. Where unaccounted mortalities trend, however, and if they are of sufficient scale, problems can occur. If depradation is greater in specific areas and mortalities are allocated by area, as is the case for Pacific halibut, then the unaccounted mortality could become very important. Given experience from the FISS, working with commercial fishers in areas susceptible to whale depradation to quantify possible losses would appear to be feasible. Some simple 'what if' model runs with assumed trends in the scale and nature of depradation could be made quite quickly as part of the 2019 SA or, more pertinently, Management Strategy Evaluation (MSE) processes to gauge what level of depradation might be important (see *Research priorities, Data related* research item 9).

Pacific halibut are landed gutted and the sex ratio of the commercial catch has therefore not been monitored historically. As the fishery is highly size selective and males and females have different growth schedules, the commercial sex ratio is not expected to be 50:50 and could vary spatially and/or temporally. As reported in Stewart and Hicks (2019), this has been a cause for concern in the SA for some years. The current IPHC 5-year Biological and Ecosystem Science Research Plan for 2017-2021 recognises the need for accurate sex identification of commercial landings both for SA and MSE work (see: https://www.iphc.int/uploads/pdf/besrp/2019/iphc-2019-besrp-5yp.pdf). In line with the plan, port-based fin clip processing was carried out during 2017 and 2018 with genotyping of samples to determine sex also conducted. The work has yet to be published but is outlined briefly in https://www.iphc.int/uploads/pdf/priphc/priphc02/ppt/iphc-2019-priphc02-05c-p.pdf. To date, the 2017 samples have been genotyped and results made available for the 2019 SA development work. The results are briefly outlined in Stewart and Hicks (2019) and are used in the 2019 individual model bridging exercise (see below). The 2017 data became available in February 2019 and it is unclear if the 2018 sex ratio results will be available for the final 2019 SA or only in 2020 for the 2021 update.

Including coastwide and regional sex ratio information in the SA is clearly important given the nature of the fishery and potential implications for model fitting (see below) and management. The willingness of IPHC to pursue important data collection and use new data in analyses is commendable. The research plan currently only includes fin clip collection in 2017 and 2018. It

may be necessary to update the plan to monitor in future years as well in case of temporal or spatial changes in sex ratios, with potentially serious implications for SA modelling. If the 2018 results are similar to the 2017 ones then the final 2019 SA may remain appropriate and credible but if the 2018 results become available in early 2020 and show different patterns, it could undermine confidence in the2019 SA and any decisions made by the Commission in January 2020. Ideally, the 2018 results would be available for the final 2019 SA.

Fishery independent information is available through the IPHC FISS and the NMFS trawl survey in Alaska. It is unusual for SA purposes to have access to even one high quality fishery-independent index and the IPHC is fortunate to have two, with the dedicated IPHC FISS being exceptional by any standard. Its duration, scope and fine-scale provide a fishery independent index (coastwide or by region or area), composition data, and biological information, including annual estimates of stock distribution by area. The FISS provides the primary index for the SA. As an IPHC-run annual survey it also provides a platform for other research (see, e.g.: https://www.iphc.int/uploads/pdf/priphc/priphc02/ppt/iphc-2019-priphc02-05b-p.pdf). The use of multiple commercial vessels further provides an opportunity for industry and Secretariat interaction and for building credibility in any outputs from the survey as used in SA. Expansion work in the FISS from 2014 through 2019 demonstrates both a flexibility seldom seen in more general surveys and a desire to improve information and credible science support for decision-making. Critically, the FISS provides information to the SA on fish below the commercial MLS of 32". Together with the NMFS survey which samples still smaller/younger fish, the FISS is a key component of the SA and provides the ability to provide probabilistic forecasts of the impacts of future catches on stock status.

The FISS is simply but well described in Webster (2019). Since 1998, it has been *undertaken annually using a 10 nmi fixed grid design, within depths of 37-503 m (20-275 ftm). This design ensures that, on average, all habitat types within the area covered by the setline survey are sampled in proportion to their occurrence, while fishing the same fixed stations each year reduces uncertainty in any estimates of trends in density indices derived from the setline survey data.* As reported in Webster (2019), the FISS has been analysed using a space-time modelling approach since 2016 but, as commented on by the SRB (IPHC, 2018): *NOTING that this is the sixth review of the spacetime modelling approach, the SRB reiterated its ENDORSEMENT of the approach as cutting-edge and could be widely used. Thus there is a pressing need to publish the space-time modelling approach used for the fishery-independent setline survey data in a peer-reviewed scientific journal.* I have been unable to find even a source grey paper on the IPHC space-time modelling, only on results and discussions such as Webster (2019), but agree with the SRB as to the general utility of the approach which is now becoming commonplace as a replacement for design-based modelling and is well understood (see, e.g.: http://www.capamresearch.org/Spatio-Temporal-Modelling-Mini-Workshop/presentations). The approach allows not just surface fitting for integration of indices but a deeper exploration of covariates and time-dependencies than more traditional approaches, as well, potentially, of

estimating biological data such as age compositions. This is commented on under *Research priorities, Data relates issues* item 12.

The SRB (IPHC, 2019a) has requested: *analysis of past prediction patterns (a type of cross-validation analysis) to help assess the proposed methods' ability to meet precision targets while maintaining low bias. This should include an examination of spatio-temporal residual patterns for the appropriateness of estimated autocorrelation*. SRB reports are summary documents and do not provide documentation of discussions leading to request (though full audio recording is available). I am therefore unclear as to the reason for the SRB request. As I understand it, it is not requesting cross-validation *per se* but the requested work is regarded as conceptually related to cross-validation. Clearly, it relates to estimates from the space-time modelling and their use in the SA. I have what might be a related comment motivated by use of the space-time modelling to understand fundamentally how the distribution of fish is more or less stable through time and how complex, and the factors that influence variation. Fixed station design will generally reduce variance but at the possible expense of bias, especially if the complex distribution of fish changes through time. The space-time modelling approach used for FISS analysis can account for variations in distribution but bias will still depend on survey coverage compared to stock distribution. The expansion work since 2014 (one area *per* year) is clearly aimed at re-design to reduce bias in estimates by area and also further reducing the variance of estimates. Any re-design of the FISS following completion of the expansion series should be beneficial.

Consideration of covariates (e.g., Dissolved oxygen) in the space-time analyses appears to be ongoing and discussion between the Secretariat science staff and the SRB is guiding inclusion or otherwise. I see no need to add further comment other than the process is working, discussions taking place, and results being produced as required for the SA.

Primary and even grey literature on the FISS and application of space-time models is scarce; it would be good to see a publication not just on methods applied to the FISS and utility in SA, but also on fundamental understanding of halibut.

One issue of note regarding FISS indices is as outlined by the SRB (IPHC, 2018) - the need for re-evaluation of whale depredation accounting in the FISS time-series. This is effectively handled in the bridging exercise (see below) using revised FISS indices estimated using data revised due to redefined and reviewed criteria for determining when a FISS station has experienced whale depradation and should therefore be deemed ineffective. The details of the revised FISS indices are not given in Webster (2019) or Stewart and Webster (2019) as the work was only completed in February 2019. Presumably they will be included in the update paper dated 2020. The issue is briefly described in Stewart and Hicks (2019). This is mentioned here primarily to emphasise that the IPHC is responsive to concerns and through iteration with the SRB is careful to address issues - in this case, requiring a revision of data usage in analyses of the FISS, re-running of the FISS and consideration within the SA development phase.

While the commercial fishery samples fish from 32” upwards, mostly age 8 upwards, the FISS samples fish from 4-5 years old and the NMFS trawl survey samples fish from 2 years old. Sampling from all sources is clearly variable but IPHC samplers are involved in both surveys as well as at ports. Age composition data are available from all sources and information on cohort structures appears coherent between sources and informative in the SA. Work on age-determination has been ongoing and current ageing appears to be robust.

The overwhelming issue that stands out from biological sampling in the FISS, NMFS Alaska survey, and commercial landings is the strong trends in weight-at-age. While not discussed in Stewart and Hicks (2019) or Stewart and Webster (2019) the issue is included under *Research priorities, Biological understanding* item 4 and PHC-besrp, 2019 already (Appendices II and III) includes a number of growth-related studies due to feed in to the SA and MSE. It is unclear at this proposed item what additional work, if any, is envisaged. As a general comment, distinguishing between the range of factors listed (*competition, density dependence, environmental effects, size-selective fishing and other factors*) is likely to be extremely difficult in practice, even with the extensive and high quality data available on Pacific halibut, other stocks, and the environment from the USA and Canada NW and USA Alaska regions. Also, while understanding historic variations in growth in relation to a number of factors might be possible, prediction is only possible if the processes are understood. As reference points are defined as spawning biomass relative to dynamic, unfished spawning biomass, changes in weight-at-age are masked in advice on Stock Status but do, of course, flow through to Decision Tables as absolute values of Total Mortality used, as well as to Trend assessments. In the case of advice on Stock and Fishery Trends apparent risks are potentially confounded and probabilities poorly determined in weight-at-age trends are not appropriately predicted. For the 3 year forecasts used this may not be problematic but is something that might be considered in the MSE.

Individual Model Development

ToR bullet 2: Aspects of individual model development. [Aspects of developing individual models to consider for including in the ensemble.]

Stewart and Hicks (2019) describes clearly the historical development of individual models given the history of fisheries, data, survey developments, problems with previous models, etc. The rationales for model development and current selection within the ensemble are well-made and I see little need to revise these core models which have been used to provide advice for a number of years. The issue of whether they might be considered separately in providing multi-model advice or using an ensemble is a separate issue considered below. Each individual model is structurally distinct and is fitted to different data, allowing an exploration of model uncertainty.The models use either the long or short time-series and for each use more (AAF) or less (CW) disaggregated abundance and composition data. Models also differ in assumptions about selectivity, natural mortality, and other factors, with time-varying selectivity in the AAF models a major feature. The Long models also incorporate a simple environmental regime

factor, coded as a binary PDO productivity regime parameter in the stock-recruit relationship and consistent with Pacific halibut SA practice over more than a decade. Further comment on the PDO is made at *Research priorities, Technical development* item 9. As noted above, the information between data sets is reasonably coherent - abundance indices are apparently correlated, despite even sex ratio differences between surveys and commercial fisheries, and, as modeled, composition data provide reasonable information on selectivity and natural mortality sufficient to allow coherent interpretations within models. I note the use of direct weight-at-age data coupled with time-varying selectivity in the AAF models; while highly parameterised it is not statistically over-parameterised. The rationale provided that the approach deals effectively with historic retrospective patterns is reasonably convincing, though there do appear to be recalcitrant retrospective patterns still associated with male selectivity estimation.

While the abundance indices provide a robust definition of scale, the greatest uncertainty is of course due to process misspecification of natural mortality, selectivity, and recruitment but the 4 models capture a wide range of that misspecification. Despite the rigorous approach to tuning, Stewart and Hicks also downweight composition data relative to abundance data which provide information on scale critical to the risk assessment.

For the current tuning approach, clearly described in Stewart and Hicks (2019; pp. 27-29) it would be useful diagnostically, even with a simple 2x2 ensemble, to track the weights applied to each of the data sources for individual models, from assessment to assessment. It is noticeable, for example, in Stewart and Hicks (2019, Fig 13) that the AAF Long tuned model estimates of trend are markedly different to the 2018 corresponding model (at least pre-1995), perhaps implying different weighting, though other individual models within the ensemble are all similar. With no simple comparison of outputs through time (e.g., such as a 2018 equivalent of Stewart and Hicks, 2019, Fig. 62) or of final tunings (Table 11), it is hard to determine the degree to which tuning per se might be an issue. This links below to *Research priorities, Technical development* item 2. Of course, as decision-making is determined by post-1995 estimates and as trigger reference points are approached increasingly by ensemble lower/mid tail estimation, the AAF Long model may not in any case be as important as either coastwide model which have lower spawning biomass scales. With the full 2019 data yet to be used in the assessment and final tuning still to be carried out, this will all change and it is not necessary to dig too deeply at this stage.

While not made explicit in Stewart and Hicks (2019), for each model, the bridging analyses presented suggest a consistent weighting and tuning of data with past corresponding model implementations, except perhaps in the case of the AAF Long model. From the report, it is unclear to what extent individual model relative weights and tuned effective weights may have changed between years. In discussion, however, it has been clarified that within-model data weighting has been kept constant year-to-year to reduce/avoid changes to model structure during annual updates. The explanation for the clear difference in estimated trends for the 2019 AAF Long model is thus that the re-tuned weighting "*was 'catching up' with all the new information added since 2015*". This is sensible practice, consistent with the approach of annual

updates. Annual updating of data includes not just newly acquired data but also re-worked data and it could be argued that even annual updates should involve complete re-weighting and re-tuning; however, re-weighting would hide effective changes in model structure. Nevertheless, for the final SA, it might be useful to see how relative weights within individual model fits might have changed through time.

There are still axes of uncertainty such as steepness which is fixed in all individual models though has already been explored to a degree. The SRB (2019a) has requested a coarse profile of steepness. Comment is made on this in the section below on the ensemble as well as in *Research priorities, Technical development* item 2. Overall, given the historical rationale and data availability, the 4 models as structured, provide a sound basis for the risk assessment provided as advice to the Commission. None of the models is regarded as right or good enough to provide advice in isolation but the set appears to capture wide structural uncertainty and the models jointly have utility. Stewart and Hicks (2019) reports on attempts to estimate steepness. There appears to be little information to allow estimation of steepness which is, of course, confounded with natural mortality and influenced in fitting by other parameter choices. Likelihood profiling on steepness will be interesting but models that can trade steepness for other parameters generally will have little impact on probabilistic advice. However, the CW Long model is the lowest scaled of the 4 models and the one for which steepness estimation to date does have an apparent impact. Any profiling will need careful tuning but should it lead to use of a steepness axis for any or all of the 4 models in the ensemble, perhaps nested weighting could be applied such that while the four structurally different models are each weighted equally, weighting within models across the additional axes (steepness) might rely on standard approaches such as AICc (Suguira, 1978).

There is one area of potential concern. The issue of stock structure and migrations is clearly recognised by the IPHC science teams, both within the existing stock boundaries of the SA but also, potentially, as pertains to connection to the western Pacific. I note in Stewart and Hicks (2019) there is just one passing reference, in *Other Uncertainty Considerations,* to the possibility of linkage to Russian waters. It receives no mention in Stewart and Webster (2019), nor in either the presentations given to the 1st session of PRIPHC02 or the current 5-year research plan. In discussion, however, the issue was raised by IPHC staff. In contrast, migration and distribution within existing stock boundaries is well-covered in the current 5-year research plan, with dedicated projects and collaborations that explore larval and early juvenile dispersal modelling, late juvenile migration using wire tags, and tail pattern recognition to follow fish through time. Stock structure and migration issues are always important and work to understand the issues is warranted. However, the existing ensemble of models includes AAF models which allow annually varying selectivity estimation. Arguably, while modelling different processes, these models should capture some of the uncertainty that might be due to migration or stock structure. The final research priority in Stewart and Hicks' list (*Research priorities, Technical development* item 9) also touches on this general issue and comment is made below. In summary here, while the issues of stock structure and migration are recognised as important to

understand, they are not regarded as critical with respect to current individual and SA modelling and the provision of robust risk assessment and advice to the Commission.

While the SA might remain focused on the 4 individual models during the full assessment and perhaps some exploration of alternatives or nesting of axes of uncertainty within models (see section on the ensemble below), the ongoing MSE work provides an opportunity for wider investigation of structural uncertainty and could be used to guide research and SA efforts in the context of what matters to decision-making.

While supporting the continued use of the 4 individual models for the 2019 full assessment, I note that Stewart and Hicks (2019) is a weigh point and that fitting to data in November 2019 could reveal issues that warrant further investigation. The initial bridging work has utilised the most recent data to address issues raised by the SRB (IPHC, 2018) regarding whale depradation in the fishery independent setline survey (FISS) and sex ratio of the commercial catch (using fin clip sampling). It is important to note that the final 2019 SA will use data up to late 2019, including from the 2019 FISS (possibly including Region 3 expansion), mortality estimates, age compositions, weights at age, and a second year of sex ratio data. Working from the weigh point, however, and the careful bridging work carried out, it appears that issues considered have either nil effect (change in software version, and consideration of whale depradation in the survey) or result in changes as expected (use of new sex ratio data).

The explanation in Stewart and Hicks (2019) of manual, iterative tuning methods used in the SA is clear and informative; far more so than most stock assessment reports. It describes well both philosophy and, to the extent possible, practice. As described and discussed during the site visit, the Pacific halibut tuning process is rigorous. Like all manual, iterative fisheries model tuning, however, it is highly time consuming, difficult to describe in complete detail, difficult to replicate, and hard to review externally given the highly detailed process.

Stewart and Hicks note the possibility of estimating observation and process error (Thorson, 2018) rather than iterative, manual tuning. Thorson outlines how recent advances in parameter estimation involving random effects could be used to replace manual tuning in fisheries assessment models. While restricting discussion to three areas of parameter tuning that might be replaced by estimation variance parameters directly, Thorson argues that the techniques are likely extendable to the case of multiple variance parameters (as required in fisheries SA such as for Paciifc halibut). It is not clear if the Pacific halibut SA could be implemented using random effects models to estimate parameter variances (in place of manual tuning) in the 2019 SA round, but it seems unlikely given the SA is currently implemented using Stock Synthesis (“SS”; Methot *et al*, 2013)) which does not yet include the option. It is well beyond the scope of this review to suggest SS might be converted to implement random effects models but Thorson notes two modelling tools that do use random effects (STAN and TMB; references in Thorson, 2018) are already available and used for stock assessment modelling. Coding the individual Pacific halibut models using STAN or TMB is a major task and unlikely within the 2019 SA round but could be explored in 2020, perhaps for comparison with updated models using manual

tuning. This is an exciting area of development that could result in a major step forward in undertaking fisheries assessment. While estimating variance parameters will be computationally time-consuming it should be much faster and 'safer' that manual, iterative tuning. Potentially, it could also be incorporated in to grid-based operating models used in MSE/MPE.

While the approach advocated by Thorson has clear advantages, it potentially has some disadvantages. One potential disadvantage is the opportunity to press a button rather than explore. The Pacific halibut SA is an excellent example of where dedicated analysts with sufficient time to focus on a stock assessment have dug deeply into data and model variants and understand individual fits. Further, a deep understanding of information content of data allows some subjective decisions to be taken; the obvious example in this (and many) cases being the priority given to abundance indices over composition data.

Stewart and Hicks (2019) point to the potential to move to Bayesian integration of the stock assessment. Advantages of using Bayesian integration are outlined in the main document: i) better characterisation of uncertainty with ii) direct interpretation of probabilities, and iii) avoiding the potential for MLE fits to mis-estimate key quantities of interest in complex models with skewed distributions. A Bayesian analysis of the CW Short model is reported in Stewart and Hicks (2019). The time taken to run the simplest of the individual models, with slightly simplified selectivity parameterisation, is of the order of two weeks. The results from the Bayesian run as only briefly reported suggest little difference to median estimates from the standard MLE run and little skewness in the Bayesian posteriors - though a hint of right skewness in male natural mortality. It is unclear if full bayesian integration of the AAF models might lead to greater differences to MLE equivalent runs but it is clear that the computing time requirements will increase and that perhaps, further simplifications will be required. From a purely practical perspective, therefore, while moving to Bayesian analyses could be done, it does not seem to be a high priority in the context of providing robust and credible decision-support. Even with the current 2x2 ensemble, Bayesian integration would be computer intensive and time consuming and could require additional time to simplify models to run efficiently. The time taken would increase as more models were potentially added to the ensemble (*Research priorities, Technical development* item 2). As indicated in the proposal, however, using Bayesian integration could provide a more natural approach for combining models in the ensemble. The current 4 individual models are all structurally different and fit to four different, though overlapping, data sets. As such, standard model weighting (AIC and BIC variants) cannot be applied regardless of MLE or Bayesian approaches being used. Alternative approaches such as Leave-One-Out cross-validation (LOO) and the Widely Applicable Information Criterion (WAIC) (see, e.g., Vehtari et al, 2017) might be applicable but would add substantially to computing time. There is no need in the current round of SA development during 2019 to investigate further Bayesian approaches but if time permits, and perhaps when the MSE work progresses and the Commission adopts simple annual catch updating mechanisms that free up SA time, further work could (as noted by Stewart and Hicks, 2019, p91) be undertaken on individual model Bayesian integration and potentially on weighting of Bayesian models in the ensemble.

Ensemble/Weighting

ToR bullet 3: The collection of models contributing to the ensemble, and the methods for combining/weighting the results.

Consideration of the ensemble needs to include i) the general methods used, including weighting of models within the ensemble; ii) preliminary results for the 2019 SA *cf* the 2018 final results; and iii) options for development.

With regard to methods (i), the approach has been developed over the past 4-5 years and is carefully explained in Stewart and Hicks (2019). Assumptions (notably the correlation between spawning biomass and the dynamic unfished spawning biomass) have been tested for impacts on key estimates used in decision-making. Provision for flexible weighting is included in the general methods. To date, individual models have received equal weighting in the ensemble as used to generate decision tables for use by the Commission though it is clear that alternatives have been explored and considered by the Secretariat and discussed with the SRB. These are noted in the section below on possible development. The approach in use is pragmatic and reasonable; it has provided the basis for a single stream of science-based risk assessment. Importantly, by using the selected ensemble of structurally different models, and not focusing on a specific model run, the Secretariat has managed largely to separate science from policy in the support materials provided to the Commission for annual decision-making. Continued use of the 2x2 ensemble as is, with equal model weighting, would continue to provide a robust and consistent approach if used in the final 2019 SA.

Stewart and Hicks (209) provide preliminary results for 2019 and compare quantities of interest estimated using the in-development SA with those made in the final 2018 SA. Usefully, Stewart and Hicks distinguish the sources of any changes in estimates. The final 2019 SA will use fully updated fishery dependent and fishery independent data sets and all individual models will be carefully re-tuned. Preliminary results therefore need to be treated with care and only potentially as aids in thinking about model development.

The preliminary SPR estimates of interest reported in Stewart and Hicks on page 87 are given in the text only and not in preliminary decision tables or any presentation I can find. This is sensible in a development document and is noted here not as a criticism but as an indication of good process; it would be dangerous to put these figures in to any other form until the final SA is completed and final decision-support material is provided. The estimates are included at this stage to enable a deconstruction of why there are changes in the estimated status compared to the 2018 SA. Understanding this is important in providing advice in a continuous decision-making context and is critical to building credibility and trust in the advice, especially if the new estimate in the final 2019 SA remains well below the 2018 estimate and close to the trigger point for the IPHC control rule. A similar deconstruction in the final SA document is encouraged.

Individual models differ in how much flexibility they assume/allow in a variety of features and only the longer time-series models use PDO data in fitting the stock-recruitment relationship. However, while the individual models are structurally different, all are fit to the same later period fishery dependent and fishery independent data in a more or less aggregated form. It is to be expected, therefore, that they will estimate the same general late period trends and with similar uncertainty, though with different assumptions or estimates of productivity translating in to different scales of spawning biomass and recruitment and hence potential yield. This appears to be the case (e.g., Stewart and Hicks, 2019; Figs. 62-64).

The change in apparent status in the 2019 preliminary SA compared to the final 2018 SA is attributed to a change in reference points, which are estimated annually as dynamic unfished SPR, updated data and "*updating of the individual models*". Changes in dynamic reference points are natural and apparently within the range of estimation as seen through Table 14 of stewart and Hicks (2019). The majority of change is attributed jointly to new data and model updates.

The key comment at this stage is that the approach to disentangling sources of change is important and useful. However, from the preliminary analyses, it is unclear to what extent individual model effective and relative weights may have changed between years using standardised approaches requiring iterative tuning. For the current tuning approach, clearly described in Stewart and Hicks (2019; pp. 27-29) it would be useful diagnostically, even with a simple 2x2 ensemble, to track the relative weights applied to each of the data sources for individual models. It is noticeable, for example, in Stewart and Hicks (2019, Fig 13) that the AAF Long tuned model estimates of trend are markedly different to the 2018 corresponding model (at least pre-1995), perhaps implying different weighting, though other individual models within the ensemble are all similar. With no simple comparison of outputs through time (e.g., such as a 2018 equivalent of Stewart and Hicks, 2019, Fig. 62) or of final tunings (Table 11), it is hard to determine the degree to which tuning *per se* might be an issue. This links below to *Research priorities, Technical development* item 2. Of course, as decision-making is determined by post-1995 estimates and as trigger reference points are approached increasingly by ensemble lower/mid tail estimation, the AAF Long model may not in any case be as important as either coastwide model. With the full 2019 data yet to be used in the assessment and final tuning still to be carried out, this will all change and it is not necessary to dig too deeply at this stage. For the final SA, it might be useful to see how relative weights within individual model fits might have changed through time.

With regard to future development (iii), the models are currently equally weighted but there is a clear concern that this might not be the most appropriate approach. Consideration needs to be given to a) weighting of the existing 2x2 ensemble, either pragmatically or formally; and b) adoption of more and/or alternative models within the ensemble. It is important to distinguish academic issues related to model weighting from weighting as it affects the quality of risk assessment provided for decision-making; i.e., Decision Tables.

The current 4 individual models in the 2x2 ensemble are all structurally different and fit to four different, though overlapping, data sets. As such, standard model weighting such as AIC and BIC variants cannot be applied regardless of the use of MLE or Bayesian approaches in individual model fitting. If Bayesian integration is progressed then alternative approaches such as Leave-one-out cross-validation (LOO) and the widely applicable information criterion (WAIC) (see, e.g., Vehtari et al, 2017) are available but would require considerable increases in both individual model computation time and in the time required for combination of those models. They are possible means of weighting that could be explored for future use if the SA adopts a Bayesian approach.

Generally, a weighted average ensemble (as used currently in the SA) is an approach that allows multiple models to contribute to a prediction in proportion to their trust or estimated performance. In the language of machine learning and neural networks this is commonly referred to as “skill”. Stewart and Hicks (2019) reports on a number of suggested weighting approaches that have been discussed in recent years with the SRB, but not progressed for reasons that are not explicit. These are to weight models in the ensemble according to i) fit to the survey index of abundance; ii) retrospective performance (using Mohn’s rho); and iii) predictive performance (i.e., skill in predicting the terminal survey index value). Ensemble weighting based on (i) places weight on models which are already likely to be more weighted to the survey in the individual model tuning phase. Weighting using retrospective performance (ii) may favour models less influenced by the treatment of male selectivity - presumably by effectively weighting to abundance *cf* composition data. Weighting based on predictive skill for the terminal survey indice (iii) is an effective, additional weight on the survey and arguably akin to selecting, or at least prioritising,composition data over indices; in that case, a more traditional approach of using different individual models separately to reveal uncertainty might be more ‘honest’. All approaches have clear rationales but the third, notwithstanding the comment above, using “skill” arguably has the best academic foundation, borrowing in concept from machine learning and neural networks. All, however, are in fact arbitrary and as individual model tunings vary through time it is likely weighting through re-tuning of models in the ensemble may also vary, hiding relative contributions to risk-based advice. Perhaps most importantly, however, all suggestions place value on fitting specific data or achieving SA stability. It would be equally plausible to suggest, for example, that in the absence of a model with explicit stock structure and movement, the AAF models should be afforded greater weight because they provide a proxy mechanism and allow for spatial and temporal variation in distribution. While all models are caricatures and our interest in them is primarily in their predictive capabilities, given the knowledge on spatial differentiation are the CW models even admissible regardless of fit diagnostics?

The IPHC has gone to great lengths to separate science from policy advice. Arguably, rather than model weighting based on fitting criteria or *a priori* “best” model consideration, weighting might instead be focused on how robust is the advice using models combined in the ensemble. All current individual models display similar trends and variances which largely affect stock status estimates equally, but they differ in estimated scale of SB and therefore potential yield

and forecasts. In decision-making that attends to probabilities of bad things happening given absolute values of catch, it is the mid lower tails of the ensemble distributions that generally might become important. The CW models have lower SB and presumably therefore lower potential yield than AAF models (e.g., Stewart and Hicks, 2019; Table 13 and Fig. 62). Therefore, even though the 4 models are currently equally weighted, for any absolute catch assumption in the decision tables based on all 4 models the estimated probability of being below stock status trigger reference points will depend on how much the CW models (with lower SB estimates) are weighted. As decision-making is concerned with the mid lower tails, the CW models have more influence on decision outcomes than the AAF models.

One easy way to evaluate the robustness of advice to weighting would be a simple, manual leave one out approach using equal weights for each combination of three models - *a priori* it might make little difference in the stock trends part of the Decision Tables though presumably would impact more on stock status 'probabilities'. Similarly, various *ad hoc* arbitrary re-weighting of the 4 models could be considered as a sensitivity test on advice.

A consistently applied and academically defensible weighting process would be ideal but the current equal weighting approach has the merit of apparent consistency and simplicity, and therefore of credibility with users. Continuing to use the approach with equal weighting is sufficient to support consistent decision-making by the Commission but investigating the robustness of the advice to different weighting, which can be done informally, would be a good first step. In the future, if SA time is freed up following use of MSE, use of a Bayesian approach, or perhaps 'automated' tuning as suggested by Thorson (2018; see also *Research priorities, Technical developments* item 3), then more formal weighting methods might be considered, explored, and used.

The use of additional or alternative individual models in the ensemble has been mooted. The SRB (IPHC, 2019a) has requested: … *Evaluate a profile (coarse) over steepness, e.g. 0.65 and 0.85, and check the impact on recruitment estimates and RSB values…* It is not clear from the SRB summary report if this request is simply aimed at further investigation of the use of a fixed value of 0.75 for steepness, or whether it is aimed possibly at *Research priorities, Technical developmen*t, item 2 and the possibility of including additional axes of uncertainty in the ensemble. Stewart and Hicks (2019) reports on attempts to estimate steepness. There appears to be little information to allow estimation of steepness which is, of course, confounded *inter alia* with natural mortality and influenced in fitting by other parameter choices. Likelihood profiling on steepness will be interesting but models that can trade steepness for other parameters generally will have little impact on probabilistic advice. However, the CW Long model is the lowest scaled of the 4 models and the one for which steepness estimation to date does have an apparent impact. Any profiling will need careful tuning but should it lead to use of a steepness axis for any or all of the 4 models in the ensemble, perhaps nested weighting could be applied such that while the four structurally different models are each weighted equally, weighting within models across the additional axes (steepness) might rely on standard approaches such as AICc (Suguira, 1978).

The ensemble has been stable for a full SA cycle (between full assessments) and provides a consistent basis for robust decision-support. While a full assessment is an opportunity to adjust individual models and the composition and/or weighting of the ensemble, any change needs to be well justified and tested for robustness. Investigating axes of uncertainty is a key part of SA but the provision of consistent, robust and credible risk assessment as a basis for regular decision-making must be considered. With MSE work currently being carried out by the IPHC and due for presentation and possible implementation in 2021, it might be prudent to minimise or even avoid any changes to the composition of the ensemble at this time.

Research Priorities

ToR bullet 4: Comments on research priorities or avenues for data, model or management advice development as appropriate.

Stewart and Hicks (2019) provide an extensive list of '*Research priorities*', spanning improvements in basic biological understanding, investigation of existing data series and collection of new information, and technical development of models and modelling approaches. The list subsumes all potential data-related future analyses highlighted by Stewart and Webster (2019). For simplicity, the complete list from Stewart and Hicks (2019) is included here as numbered items, together with comments. The text from Stewart and Hicks is in *blue italics*. Comments are in black. Potential recommendations on prioritisation are underlined and **possible priorities are in bold case**. Note that Stewart and Hicks (2019) is a complete list and does not suggest potential costs and benefits or prioritisation, nor does it distinguish work already started from work that is proposed. In the final SA report due in September 2019, it would be helpful to separate in progress from suggested future work and for suggested work to provide priority rankings with justification, ideally linked to the text of the main report. This would assist reading but would also integrate better with development and updating of 5-year plans.

NOTE: The 5-year research plan reported in Planas (2019) seems now to be replaced by https://www.iphc.int/uploads/pdf/besrp/2019/iphc-2019-besrp-5yp.pdf. I can find no formal reference to this document and it is referred to in this report as IPHC-besrp, 2019.

Biological understanding

During the last several years, the IPHC Secretariat has developed a comprehensive five-year research program (Planas 2019). The development of the research priorities has been closely tied to the needs of the stock assessment and harvest strategy policy analyses, such that the IPHC's research projects will provide data, and hopefully knowledge, about key biological and ecosystem processes that can then be incorporated directly into analyses supporting the management of Pacific halibut. Key areas for improvement in biological understanding include:

1. *The current functional maturity schedule for Pacific halibut, including fecundity-weight relationships and the presence and/or rate of skip spawning.*This is already in progress

as reported in Planas (2019), IPHC-besrp, (2019), and Stewart and Webster (2019); no further comment.

2. *The stock structure of the Pacific halibut population. Specifically, whether any geographical components (e.g., Region 4B) are isolated to a degree that modelling approximations would be improved by treating those components separately in the demographic equations and management decision-making process.* See also item 3, below.
3. *Movement rates among Biological Regions remain uncertain and likely variable over time. Long-term research to inform these rates could lead to a spatially explicit stock assessment model for future inclusion into the ensemble.* The issue of stock structure and migrations is clearly recognised by the IPHC science teams, both within the existing stock boundaries of the SA but also, potentially, as pertains to connection to the western Pacific. I note in Stewart and Hicks (2019) there is just one passing reference, in *Other Uncertainty Considerations,* to the possibility of linkage to Russian waters. It receives no mention in Stewart and Webster (2019), nor in any of the presentations given to the 1st session of PRIPHC02 or the current 5-year research plan. In discussion, however, the issue was raised by IPHC staff, consistent with general descriptions on the IPHC website (https://iphc.int/management/science-and-research/pacific-halibut-stock-status-and-biology). In contrast, migration and distribution within existing stock boundaries is well-covered in the current 5-year research plan, with dedicated projects and collaborations that explore larval and early juvenile dispersal modelling, late juvenile migration using wire tags, and tail pattern recognition to follow fish through time. Stock structure and migration issues are always important and work to understand the issues is warranted. However, the existing ensemble of models includes AAF models which allow annually varying selectivity estimation. Arguably, while modelling different processes, these models should capture some of the uncertainty that might be due to migration or stock structure. The final research priority in Stewart and Hicks' list (*Technical development*, item 9) also touches on this general issue and comment is made there. In summary here: i) while the issues of stock structure and migration are recognised as important to understand, they are not regarded as critical with respect to current individual and SA modelling and the provision of robust risk-based advice to the Commission; ii) spatial distribution and migration are already incorporated into the 5-year work program; and iii) the issue of connection between eastern and western Pacific stocks is not currently covered in IPHC-besrp, 2019, but warrants investigation and reporting in the full SA report **(Medium priority)**
4. *The relative role of potential factors underlying changes in size-at-age is not currently understood. Delineating between competition, density dependence, environmental effects, size-selective fishing and other factors could allow improved prediction of size-at age under future conditions.* IPHC-besrp, 2019 already (Appendices II and III) includes a number of growth-related studies due to feed in to the SA and MSE. it is unclear at this proposed item what additional work, if any, is envisaged. As a general comment, distinguishing between the range of factors listed is likely to be extremely difficult in practice, even with the extensive and high quality data available on Pacific halibut, other

stocks, and the environment from the USA and Canada NW and USA Alaska regions. Also, while understanding historic variations in growth in relation to a number of factors might be possible, prediction is only possible if the processes are understood. **(Unclear priority)**

5. *Improved understanding of recruitment processes and larval dynamics could lead to covariates explaining more or the residual variability about the stock-recruit relationship than is currently accounted for via the binary indicator used for the Pacific Decadal Oscillation.*This appears to be subsumed under *Technical development*, item 8.
6. *Improved understanding of discard mortality rates and the factors contributing to them may reduce potential biases in mortality estimates used for stock assessment.*This appears to be subsumed under *Data related research*, item 11.

Data related research

This section represents a list of potential projects relating specifically to existing and new data sources that could benefit the Pacific halibut stock assessment.

1. *Continued collection of sex-ratio from the commercial landings will provide valuable information for determining relative selectivity of males and females, and therefore the scale of the estimated spawning biomass, and the level of fishing intensity as measured by SPR. Potential methods for estimating historical sex-ratios from archived scales, otoliths or other samples should be pursued if possible.* Estimates of historic and future catch sex ratios are critical to credible usage of SPR in the management context. Fin clipping of fish in the ports, together with genetic analysis, has already provided a sex ratio estimate for 2017, with a 2018 estimate imminent. This is covered in the 5-year research plan. However. The plan does not explicitly include continued fin clipping/genetic work after 2018. Nor is there any provision for estimating historic sex ratios. The potential project noted by Stewart and Hicks seems to presuppose future monitoring - this might be clarified in the 5-year research plan and the final SA report. The suggestion for methods to estimate historical sex ratios, at this stage just to explore what is possible using archived samples, is important. Consideration should be given to including at least exploration of archived samples and potential for sex ratio estimation in the 5-year plan **(Exploration - high priority)**
2. *The work of Monnahan and Stewart (2015) modelling commercial fishery catch rates has been extended to include spatial effects. This could be used to provide a standardized fishery index for the recent time-series.* The reference is not alluded to in the main text of Stewart and Hicks (2019) and is not included in the reference list. It is referenced in Stewart and Webster (2019) where it is noted that: ...*A detailed exploratory analysis of the logbook standardization data and methods was completed during 2014 (Monnahan and Stewart 2015), which suggested future analyses may be able to include all logbook records in all Regulatory Areas regardless of gear type if a model-based estimator were used. However, discussions with the IPHC's Scientific Review Board did not result in a recommendation to change the simple method employed historically*...and from which the proposal appears to carry over. Without further discussion and information it is not possible to comment or suggest priority.

3. *A revised hook spacing relationship (Monnahan and Stewart 2017) will be investigated for inclusion into IPHC database processing algorithms.* This is noted as important but, as stated, seems to be a given rather than a proposal.
4. *Reevaluation of the historical length-weight relationship to determine whether recent changes in length-at-age are also accompanied by changes in weight-at-length and how this may change estimates of removals over time is ongoing.*This is noted as important but already in progress.
5. *A historical investigation on the factors influencing observed size-at-age, and ageing of additional samples from key periods and areas to support this analysis is ongoing at the IPHC.*This is noted as important but already in progress.
6. *There is the potential that trawl surveys, particularly the Bering Sea trawl survey, could provide information on recruitment strengths for Pacific halibut several years prior to currently available sources of data. Geostatistical modelling and renewed investigation of the lack of historical correlation between trawl survey abundance and subsequent abundance of Pacific halibut in the FISS and directed fisheries may be helpful for this effort.* Early indications of recruitment are clearly key to forecasting three years ahead, as done for the decision tables provided annually. Given fishery selectivity and regulations (MLS) the FISS currently contains information 3-4 years ahead of recruitment to the fishery. The NMFS survey could in principle extend this lead in by a further 2-3 years. With annual decision-making, 3-year forecasts are likely sufficient, and if MSE leads to implementation of control rules or management procedures then FISS-derived indices are likely to dominate in informing annual mortality changes. While this proposed work would be interesting and potentially useful in developing understanding of ontogenetic or environmentally-related changes in distribution of halibut, and may be worthwhile in its own right, it is not a clear priority for SA or MSE.
7. *There is a vast quantity of archived historical data that is currently inaccessible until organized, electronically entered, and formatted into the IPHC's database with appropriate meta-data. Information on historical fishery landings, effort, and age samples would provide a much clearer (and more reproducible) perception of the historical period.* No detail on historical data (as specified in this research item) or archived materials is given in Stewart and Hicks (2019) or Stewart and Webster (2019) though Stewart and Hicks does report briefly on, e.g., re-ageing of archived otolith samples. The listed avenue of research is a general comment about inaccessible, archived data and is difficult to comment on except to provide in principle support for careful cataloguing, reanalysis and use of historical data and materials (e.g., for sex ratio estimation as at *Data related research* item 1). The re-ageing reported by Forsberg and Stewart (2015) is a good example of why such materials and data are important. It is noted that the suggestion for this item is consistent with various annual reports of assessment and research activities (e.g., IPHC, 2014).
8. *Additional efforts could be made to reconstruct estimates of subsistence harvest prior to 1991.* It is unclear from Stewart and Webster (2019), from which this item carries over, what if any sources of existing data might be used to reconstruct subsistence estimates, or if the proposal is to use e.g. structured interviewing techniques to gather information.

The scale of post-1991 subsistence estimates, however, is very small compared to other sources of mortality and it is not obvious that this work should be afforded great priority from a technical perspective.

9. *NMFS observer data from the directed Pacific halibut fleet in Alaska could be evaluated for use in updating DMRs and the age-distributions for discard mortality. This may be more feasible if observer coverage is increased and if smaller vessels (< 40 feet LOA, 12.2 m) are observed in the future. Post-stratification and investigation of observed vs. unobserved fishing behavior may be required.* Discard mortality in the directed fishery is clearly an important component to quantify and age-composition data of discards potentially provides key information on recruitment and potential yields. Increased observer coverage generally and extension to smaller vessels is clearly desirable but as commented above, while improved information would be useful and could improve credibility of the SA, it is beyond the scope of this review to recommend increasing observer coverage by a member state. This research proposal is one of a number about improving or acquiring basic data but is different in that it implies a change in monitoring. As such, with considerable cost implication, clear justification with costs and benefits to support prioritisation is required. NOTE based on the main text above: One other potential unaccounted mortality in the commercial fishery is that due to whale depradation. An exploration of potential importance in risks assessments that might be caused by trends in scale and nature of this could be undertaken quickly to determine what priority might be placed on estimating depradation in commercial fisheries. Exploration using MSE that includes how unaccounted trends impact the assessment-decision-implementation loop would be preferable. **(Medium priority)**
10. *Historical bycatch length frequencies and mortality estimates need to be reanalyzed accounting for sampling rates in target fisheries and evaluating data quality over the historical period.* It is unclear if this relates also to item 7 on inaccessible data or to accessible data sets requiring new analysis; I presume the latter. IPHC (2019c) indicates recent bycatch mortality is about 15% of total mortality but visually from Stewart and Hicks (2019; Fig. 3) historical bycatch mortality may have been as much as 25% in the 1960s and approaching 50% in the late 1970s and 1980. Older fish are well represented in the early (i.e., pre-1992) bycatch compositions. It is unclear from the main Stewart and Hicks (2019) text why this specific reanalysis is 'needed' and what priority it should receive; there is no suggestion that the data as used currently in the assessment are flawed except also by implication at *Research proposal, Technical Development* item 5. Improving these data to the greatest extent possible would be welcome and might impact on historical perspectives but it is unclear how it might flow through to impact on current advice. **(Medium priority?)**
11. *There are currently no comprehensive variance estimates for the sources of mortality used in the assessment models. In some cases, variance due to sampling and perhaps even non-sampling sources could be quantified and used as inputs to the models via scaling parameters or even alternative models in the ensemble.* (See also *Biological understanding*, item 6.) It is not uncommon to use gross sensitivity tests to account for potential misspecification of mortality components, particularly of scale, and, perhaps

more importantly, trend. This could be done as part of SA sensitivity testing and/or might be incorporated into MSE robustness testing. However, it does need to be informed by data and analysis to be credible. It is unclear from the core documents available for review what precisely is envisaged under this proposal item or if priorities would be assigned by sector. Presumably, data and information on observer coverage, etc, exist and could be used to estimate variances but issues of scale and trend may often require less formal information. Issues affecting estimates will vary by sector and information on changing practices within sectors will require careful consideration. The directed fishery is the largest proportion of mortality but likely the best sampled, though issues such as conversion factors and changing practices might be relevant. Changes through time due to regulatory change and low observer coverage might be relevant in the bycatch fishery. Over more recent times, growth in variable recreational fisheries might be of importance. It would be useful to consider this proposed item in light of perceived problems and to set priorities accordingly **(Medium priority?)**.

12. *A space-time model could be used to calculate weighted FISS age-composition data. This might alleviate some of the lack of fit to existing data sets that is occurring not because of model misspecification but because of incomplete spatial coverage in the annual FISS sampling which is accounted for in the generation of the index, but not in the standardization of the composition information.* Fitting weighted age-composition data using a space-time model would be interesting and for fisheries with less extensive sampling could be highly beneficial. However, it is not clear from Stewart and Hicks (2019) reports of individual model fits why this proposed work would be of high priority for the SA. While there is incomplete spatial coverage in the FISS age sampling, it is nevertheless extensive and fits to FISS age composition data appear generally good for all models, though I note Fig. 35 and residual patterns in the AAF Short model. The expansion work should also lead to improved age compositions. I note the comments by Thorson (http://www.capamresearch.org/sites/default/files/Thorson2.pdf; slides 46 onwards) concluding i) the feasibility of estimating age compositions using space-time models; but ii) perhaps with little benefit. However, Thorson's conclusion *re* little benefit is somewhat countered by the example used that shows stock assessment outcomes when using either design or model-based age composition data; relative spawning biomass appears little affected but in the example case the absolute spawning biomass levels are very different. Given the lack of information on scale in composition data this seems strange. Exploration of a space-time model as suggested could lead to standardised composition data as suggested and is worthy of exploration, also as an alternative/backup should future sampling or ageing be compromised. **(Not essential for the SA so Low to medium priority?)**

Technical development

There are a variety of technical explorations and improvements that could benefit the stock assessment models and ensemble framework. Although larger changes, such as the new data sets and refinements to the models presented in this document, naturally fit into the period full

assessment analyses, incremental changes may be possible during updated assessments when and if new data or methods become available. Specifically, development is intended to occur in time for initial SRB review (generally in June), with only refinements made for final review (October), such that untested approaches are not being implemented during the annual stock assessment itself. Technical research priorities include: This preamble suggests the list contains technical developments that '*could*' benefit the individual SA and ensemble but the final sentence uses the word '*priorities*'. If the intention is to prioritise then further justification is required at each item with respect to the SA and perhaps MSE but especially in the context of providing robust, consistently-based, and credible decision-support.

1. *Maintaining consistency and coordination between MSE, and stock assessment data, modelling and methodology.* Noted and supported; presumably this is ongoing and standard operating procedure. It is unclear why that this needs to be given specific mention as a "*technical exploration and improvemen*t".
2. *Continued refinement of the ensemble of models used in the stock assessment. This may include investigation of alternative approaches to modelling selectivity that would reduce relative downweighting of certain data sources (see section above), evaluation of additional axis of uncertainty (e.g., steepness, as explored above), or others.* Stewart and Hicks (2019) reports on attempts to estimate steepness. There appears to be little information to allow estimation of steepness which is, of course, confounded with natural mortality and influenced in fitting by other parameter choices. Likelihood profiling on steepness will be interesting but models that can trade steepness for other parameters generally will have little impact on probabilistic advice. However, the CW Long model is the lowest scaled of the 4 models and the one for which steepness estimation to date does have an apparent impact. Any profiling will need careful tuning but should it lead to use of a steepness axis for any or all of the 4 models in the ensemble, perhaps nested weighting could be applied such that while the four structurally different models are each weighted equally, weighting within models across the additional axes (steepness) might rely on standard approaches such as AICc (Suguira, 1978). // The ensemble has been stable for a full SA cycle (between full assessments) and provides a consistent basis for robust decision-support. While a full assessment is an opportunity to adjust individual models and the composition and/or weighting of the ensemble, any change needs to be well justified and tested for robustness. Investigating axes of uncertainty is a key part of SA but the provision of consistent, robust and credible risk assessment as a basis for regular decision-making must be considered. With MSE work currently being carried out by the IPHC and due for presentation and possible implementation in 2021, it might be prudent to minimise or even avoid any changes to the composition of the ensemble at this time.
3. *Evaluation of estimating (Thorson 2018) rather than tuning (Francis 2011; Francis 2016) the level of observation and process error in order to achieve internal consistency and better propagate uncertainty within each individual assessment model. This could include the 2d Autoregressive smoother for selectivity, the Dirichlet multinomial, and other features now implemented in stock synthesis (Methot et al. 2019).*The explanation in Stewart and Hicks (2019) of manual tuning methods/approaches used in the SA is

clear and informative; far more so than most stock assessment reports. As described and discussed during the site visit the Pacific halibut tuning process is rigorous. Like all fisheries model tuning, however, it is highly time consuming, difficult to describe in detail, difficult to replicate, and very hard to review. Stewart and Hicks note the possibility of estimating observation and process error (Thorson, 2018) rather than iterative, manual tuning. Thorson outlines how recent advances in parameter estimation involving random effects could be used to replace manual tuning in fisheries assessment models. While restricting discussion to three areas of parameter tuning that might be replaced by estimation variance parameters directly, Thorson argues that the techniques are likely extendable to the case of multiple variance parameters (as required in fisheries SA such as Paciifc halibut). It is not clear if the Pacific halibut SA could be implemented using random effects models to estimate parameter variances (in place of manual tuning) in the 2019 SA round, but it seems unlikely given the SA is currently implemented using Stock Synthesis (Methot *et al*, 2013)) which does not yet include the option. It is well beyond the scope of this review to suggest SS might be converted to implement random effects models but Thorson notes two modelling tools that do use random effects (STAN and TMB; references in Thorson, 2018) already available and used for stock assessment modelling. **Coding the individual Pacific halibut models using STAN or TMB is a major task and unlikely within the 2019 SA round but could be explored in 2020, perhaps for comparison with updated models using manual tuning.** This is an exciting area of development that could result in a major step forward in undertaking fisheries assessment. While estimating variance parameters will be computationally time-consuming it should be much faster and 'safer' that manual, iterative tuning. Potentially, it could also be incorporated into grid-based operating models used in MSE/MPE.

4. *Continued development of weighting approaches for models included in the ensemble, potentially including fit to the survey index of abundance, retrospective, and predictive performance (see section above).* As noted at item 6, below, the current 4 individual models are all structurally different and fit to four different, though overlapping, data sets. As such, standard model weighting (AIC and BIC variants) cannot be applied regardless of MLE or Bayesian approaches being used. Alternative (effectively cross-validation) approaches are available for Bayesian models (see, e.g. Vehtari et al, 2017) but would require considerable increases in both individual model computation time and in the combination of those models. They are possible means of weighting that could be explored for future use if the SA adopts a Bayesian approach. Generally, A weighted average ensemble is an approach that allows multiple models to contribute to a prediction in proportion to their trust or estimated performance. Stewart and Hicks (2019) reports on a number of suggested weighting approaches that have been discussed with the SRB but not progressed. These are to weight models in the ensemble according to i) fit to the survey index of abundance; ii) retrospective performance (using Mohn's rho); and iii) predictive performance (i.e., skill in predicting the terminal survey index value). Ensemble weighting based on (i) places weight on models which are already likely to be more weighted to the survey in the individual model tuning phase. Weighting using

retrospective performance (ii) may favour models less influenced by the treatment of male selectivity - presumably by effectively weighting to abundance *cf* composition data. Weighting based on predictive skill for the terminal survey indice (iii) is an effective, additional weight on the survey and arguably akin to selecting, or at least prioritising composition data over indices; in that case, a more traditional approach of using different individual models separately to reveal uncertainty might be more 'honest'. All approaches have clear rationales but the third, notwithstanding the comment above, using "skill" arguably has the best academic foundation, borrowing in concept from machine learning and neural networks. All, however, are in fact arbitrary and as individual model tunings vary through time it is likely weighting through re-tuning of models in the ensemble may also vary, hiding relative contributions to risk-based advice. The IPHC has gone to great lengths to separate science from policy advice; care is needed in investigating any *ad hoc* weighting to focus not on which models make a difference but on how robust is the advice using those four models. All models display similar trends and variances which affect status determination and forecasts but they differ in estimated scale of SB and therefore potential yield. In decision-making that attends to probabilities of bad things happening, it is the mid lower tails of the distributions of absolute values that generally might become important, with the CW models having lower SB and presumably therefore potential yield than AAF models (e.g., Stewart and Hicks, 2019; Table 13 and Fig. 62). One simple way to evaluating the robustness of advice to weighting would be a simple, manual leave one out approach using equal weights for each combination of three models - *a priori* it might make little difference in the status trends part and perhaps stock trends part of the Decision Tables though presumably would impact more fishery trend 'probabilities'. Similarly, an ad hoc arbitrary re-weighting of the 4 models could be considered as a sensitivity test on advice. A consistently applied and academically defensible weighting process would be ideal but the current approach has the merit of consistency and simplicity. Continuing to use the approach with equal weighting is sufficient to support decision-making by the Commission but investigating the robustness of the advice to different weighting, which can be done informally, would be a useful step in the 2019 SA **(SA 2019; Medium priority).** In time, if SA time is freed up following use of MSE, and if the SA adopts a Bayesian approach, more formal weighting methods might be used **(Post MSE)**

5. *Exploration of methods for better including uncertainty in discard mortality and bycatch estimates in the assessment (now evaluated only via alternative mortality projection tables or model sensitivity tests) in order to better include these sources uncertainty in the decision table. These could include explicit discard/retention relationships, including uncertainty in discard mortality rates, and allow for some uncertainty directly in the magnitude of mortality for these sources.* See also *Research proposals, Data related research* item 10. Work under the data related research needs to proceed first to identify uncertainties in the mortality estimates. Depending on estimates, SA and MSE focus can then be directed appropriately if warranted. The standard approach of conducting sensitivity tests on the individual models and perhaps decision tables is the obvious first approach within the SA. Including discard/retention relationships in the SA would need to

be informed by data, potentially from compliance authorities. MSE can be used to test the implications of different relationships in combination with management. If biases are consistent then the implications for decision-making are likely to be small or insignificant. If biases are variable but reasonably symmetric then the effectiveness of any control rule or management procedure will depend on its inputs (likely from the FISS) and their ability to track changes in recruited biomass. If, however, there is a discard/retention relationship related, e.g., to regulatory 'bite' (such as reducing catch limits) then unless control rules or management procedures react quickly to informative inputs, there is potential for unseen stock decline. If analyses suggest biases and especially any discard/retention relationships then the MSE rather than the SA would be an appropriate mechanism to investigate implications and to develop robust management responses as part of control rules or management procedures. **(Priority in MSE depends on analyses to identify potential issues)**

6. *Bayesian methods for fully integrating parameter uncertainty may provide improved uncertainty estimates within the models contributing to the assessment, and a more natural approach for combining the individual models in the ensemble (see section above).* Advantages of using Bayesian integration are outlined in the main document: i) better characterisation of uncertainty with ii) direct interpretation of probabilities, and iii) avoiding the potential for MLE fits to mis-estimate key quantities of interest in complex models with skewed distributions. A Bayesian analysis of the CW Short model is reported in Stewart and Hicks (2019). The time taken to run the simplest of the individual models, with slightly simplified selectivity parameterisation, is of the order of two weeks. The results from the Bayesian run as only briefly reported suggest little difference to median estimates from the standard MLE run and little skewness in the Bayesian posteriors - though a hint of right skewness in male natural mortality. It is unclear if full Bayesian integration of the AAF models might lead to greater differences to MLE equivalent runs but it is clear that the computing time requirements will increase and that, perhaps, further simplifications will be required. From a purely practical perspective, therefore, while moving to Bayesian analyses could be done, it does not seem to be a high priority in the context of providing robust and credible decision-support. Even with the current 2x2 ensemble, Bayesian integration would be computer intensive and time consuming and could require additional time to simplify models to run efficiently. The time taken would increase as more models were potentially added to the ensemble (Technical development, item 2). As indicated in the proposal, however, using Bayesian integration could provide a more natural approach for combining models in the ensemble. The current 4 individual models are all structurally different and fit to four different, though overlapping, data sets. As such, standard model weighting (AIC and BIC variants) cannot be applied regardless of MLE or Bayesian approaches being used. Alternative approaches such as Leave-one-out cross-validation (LOO) and the widely applicable information criterion (WAIC) (see, e.g., Vehtari et al, 2017) might be applicable but would add substantially to computing time. There is no need in the current round of SA development during 2019 to investigate further Bayesian approaches but if time permits, and perhaps when the MSE work progresses and the Commission adopts

simple annual catch updating mechanisms that free up SA time, further work could (as noted by Stewart and Hicks, 2019, p91) be undertaken on individual model Bayesian integration and potentially on weighting of Bayesian models in the ensemble. **(Post MSE)**

7. *Exploration of stock synthesis features previously unavailable or unevaluated including: timing of fishery and survey observations, the fishing mortality approximation used (i.e., estimated parameters, 'hybrid' or Pope's approximations).* Stewart and Hicks (2019) describe the standard population structuring adopted for all models in the SA, using mid year removals and Pope's approximation.For Pacific halibut, while exploration of alternatives may be interesting it would seem a low priority given the approximations are robust except at high fishing mortality - which is not the case. It is unclear why the proposal is made.
8. *An analysis of model sensitivity and statistical performance of treating the environmental relationship between recruitment and the PDO as annual deviates (+/-), a running mean, or annual values (actual PDO), or other methods that differ from the binary indicator variable currently employed.*The current binary indicator approach requires only a single parameter estimate (of β) in each of the Long models, and is informed primarily for the later part of the time series for which good composition data are available. It effectively assumes an unspecified linkage between general environmental state and Pacific halibut recruitment. Any alternative using e.g. a running mean or actual values in essence assumes a more direct link between PDO state and the scale of Pacific halibut recruitment resulting from the within-species contest competition implied by the Beverton-Holt S-R function. Pacific halibut recruitment, however, derives from complex and stochastic environmental processes and from complex single and multi species biological and ecological processes, also subject to stochasticity. Any direct link between PDO and recruitment will therefore have high process error, as well as observation error in the composition data informing recruitment estimation. Tuning will need to pay attention directly to recruitment but also to aliasing estimates of natural mortality in particular, but also selectivity. This would be compounded if steepness were also estimated or alternative steepness values assumed. While exploring alternative PDO linkage functions would be an interesting research area and might potentially result in apparently improved stock assessment(s) at any point in time, it is not at all clear that this would benefit risk assessments derived using stock assessments because without understanding the complex processes linking the PDO specifically to Pacific halibut recruitment, forecasting utility would not necessarily be enhanced. The MSE might again be the best place to explore how changes in environment (in a wide sense, to include not just e.g. PDO but also e.g. other species stock distribution and abundance) might affect recruitment and how alternative control rules or management procedures might be more or less robust. **(SA: Low priority; MSE: Medium priority?)**
9. *Alternative model structures, including a growth-explicit statistical catch-at-age approach and a spatially explicit approach may provide avenues for future exploration. Efforts to develop these approaches thus far have been challenging due to the technical complexity and data requirements of both. Previous reviews have indicated that such*

efforts may be more tractable in the context of operating models for the MSE, where conditioning to historical data may be much more easily achieved than fully fitting an assessment model to all data sources for use in tactical management decision making. (See also *Research priorities, Biological Understanding* items 2 and 3). The SA and MSE "philosophies" are different with more care typically taken in development of individual SA models. Conditioning, however, still requires fitting, though it is impractical to fit with the rigour used, e.g., in the individual IPHC stock assessments, especially when grid approaches with wide parameter spaces are used and specific parameter combinations may be infeasible or not well supported. Nevertheless, development of spatially explicit models for MSE purposes needs to start with careful model development and fitting as used for the tactical SA, even if final generating (operating) models are less rigorously fit. Regardless, so long as the tactical SA ensemble approach reasonably captures uncertainties through proxies for explicit spatial models (e.g. AAF with annual variation in selectivity) then specific consideration of spatially explicit models is best left to MSE where assessment and management robustness can be explored more thoroughly.

References

- Forsberg, J.E., and Stewart, I.J. 2015. Re-ageing of archived otoliths from the 1920s to the 1990s, IPHC Report of Assessment and Research Activities 2014. p. 405-428.
- IPHC (2014) Report of Assessment and Research Activities 2014 IPHC-2014-RARA24
- IPHC (2017) Report of the 93rd Session of the IPHC Annual Meeting (AM093) IPHC–2017–AM09–R3
- IPHC (2018) Report of the 13th Session of the IPHC Scientific Review Board (SRB013) IPHC–2018–SRB013–R
- IPHC (2019a) Report of the 14th Session of the IPHC Scientific Review Board (SRB014) IPHC–2019–SRB014–R
- IPHC (2019b) Summary of the Pacific halibut data, and assessment, and mortality projections IPHC-2019-AM095-08/09
- IPHC (2019c) Fishery Statistics IPHC-2019-PRIPHC02-05a
- Methot, R.D., Wetzel, C.R., and Taylor, I.G. (2019). Stock Synthesis User Manual Version 3.30.13. NOAA Fisheries. Seattle, WA. 213 p.
- Monnahan, C.C., and Stewart, I.J. 2015. Evaluation of commercial logbook records: 1991- 2013. IPHC Report of Assessment and Research Activities 2014. p. 213-220.
- Planas, J. (2019). IPHC 5-year biological and ecosystem sciences research program update. IPHC-2019-AM095-14. 7 p.
- Stewart, I.J., and Hicks, A.C. (2019). 2019 Pacific Halibut (*Hippoglossus stenolepsis*) stock assessment: Development IPHC-2019-SRB014-07. 100 p
- Stewart, I., and Webster, R. (2019). Overview of data sources for the Pacific halibut stock assessment, harvest policy, and related analyses. IPHC-2019-AM095-08. 76 p

- Stewart, I.J. and Monnahan, C.C. (2017) Implications of process error in selectivity for approaches to weighting compositional data in fisheries stock assessments. Fisheries Research 192, 126-134.
- Stewart, I.J. and Martell, S.J.D. (2015) Reconciling stock assessment paradigms to better inform fisheries management. ICES Journal of Marine Science 72(8), 2187-2196.
- Stewart, I.J. and Hicks, A.C. (2018) Interannual stability from ensemble modelling. Canadian Journal of Fisheries and Aquatic Sciences 75, 2109-2113.
- Sugiura, N. (1978). Further analysis of the data by Akaike's criterion and the finite 21 correction. Communications in Statistics, Theory and Methods A7, 13-16.
- Thorson, J.T. (2018) Perspective: Let's simplify stock assessment by replacing tuning algorithms with statistics. Fisheries Research 217:133-139
- Vehtari, A., A. Gelman and J. Gabry (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing, 2017, Volume 27, Issue 5, pp 1413-1432
- Webster (2019) Space-time modelling of survey data IPHC-2019-AM095-07