Center for Independent Experts Independent Peer Review Report

Stock Assessment Review (STAR) Panel 1 – Virtual. Dover Sole and Pacific Spiny Dogfish

May 3-7, 2021

Dr. Noel Cadigan

Centre for Fisheries Ecosystems Research
Marine Institute of Memorial University of Newfoundland
St. John's, NL. Canada

Executive Summary

The assessments for Dover Sole and Pacific Spiny Dogfish constitute the best available scientific information on the current status of the stocks. These assessments provide a suitable basis for management decisions. The SS3 stock assessment models were competently applied, and the model inputs were derived using best practice.

The major axis of uncertainty for Dover Sole was based on the final year spawning biomass. Values for female M were chosen so that model estimates of final year spawning output matched the 12.5% and 87.5% quantiles of estimates. The major axis of uncertainty for Pacific Spiny Dogfish was the WCGBTS catchability (q). Models with $q = 0.9$ and $q = 0.3$ were used as the low and high states of nature, respectively.

Stock boundaries were based on pragmatic considerations and data availability. There may be more sub-stock structure for Dover Sole than is reflected by the current assessment region boundaries. This stock appears to be fairly sedentary as adults so localized depletion could be a problem. Spiny dogfish is a transboundary stock, and there are high densities of dogfish close to the U.S.-Canada border. The U.S. and Canada should explore the possibility of a joint stock assessment in future years.

The accuracy of estimates of landings and discards has improved over time. However, there is uncertainty in catch estimates, and more so for historic periods, that is not measured and incorporated into the assessments. The accuracy of discard estimates for Pacific Spiny Dogfish is uncertain.

The most important deficiency of the Dover Sole assessment was the incomplete knowledge of stock structure and spatial productivity dynamics. The stock is not completely available to all fisheries and surveys, but the magnitude of this is uncertain. This produces uncertainty about the veracity of the cryptic biomass estimated for this stock. This also produces complicated and domed selectivity patterns which reduces information about M.

The most important deficiencies of the Pacific Spiny Dogfish assessment are uncertain catch statistics, relatively short time-series of length and age compositions with high between year variations in length, and large uncertainties in the aging of older, larger females. This species does not have high variation in recruitment, which will also tend to obscure cohort dynamics based on length and age composition data.

Background

The virtual Stock Assessment Review (STAR) Panel for Dover Sole and Pacific Spiny Dogfish was held during May 3-7, 2021. The general goals and objectives of the groundfish STAR process are to:

1) ensure that stock assessments represent the best available scientific information and facilitate the use of this information by the Council to adopt OFLs, ABCs, ACLs, (HGs), and ACTs;

2) meet the mandates of the Magnuson-Stevens Fisheries Conservation and Management Act (MSA) and other legal requirements;

3) follow a detailed calendar and fulfill explicit responsibilities for all participants to produce required reports and outcomes;

4) provide an independent external review of stock assessments;

5) increase understanding and acceptance of stock assessments and peer reviews by all members of the Council family;

6) identify research needed to improve assessments, reviews, and fishery management in the future; and

7) use assessment and review resources effectively and efficiently.

Benchmark stock assessments were conducted and reviewed for Dover Sole and Pacific Spiny Dogfish. These stocks were identified within the top five rankings for assessment consideration during the Pacific coast groundfish regional stock assessment prioritization process, which was based on the national stock assessment prioritization framework (http://www.st.nmfs.noaa.gov/Assets/stock/documents/PrioritizingFishStockAssessments_FinalWeb.pdf).

Review Panel (RP) membership is described in Appendix 3. The support of all these scientists and staff to the STAR RP process is gratefully acknowledged.

CIE reviewers were tasked with conducting impartial and independent peer reviews in accordance with their SoW and ToRs. The reviewers were required to be active and engaged participants throughout panel discussions and able to voice concerns, suggestions, and improvements, while respectfully interacting with other review panel members, advisors, and stock assessment technical teams. The CIE reviewers were required to have excellent communication skills in addition to working knowledge and recent experience in fish population dynamics; with experience in the integrated-analysis modeling approach, using age- and size-(and possibly spatially-) structured models, and methods for quantifying uncertainty. Familiarity with environmental, ecosystem and climatic effects on population dynamics and distribution may also be beneficial. The CIE reviewer's duties shall not exceed a maximum of 14 days to complete all work tasks of the peer review.

Role of reviewer

All assessment documents and most supporting materials were made available to the RP via an ftp server two weeks before the meeting, on April 19, 2021. These documents are listed in

Appendix 1. I reviewed the background documents I was provided and compiled a list of issues to get clarification at the RP meeting. I attended the entire STAR Panel review meeting via the RingCentral platform during May 3-7, 2021. I reviewed presentations and reports and participated in the discussion of these documents, in accordance with the SoW and ToRs (see Appendix 2). I drafted text for the RP report, and lead the Dover Sole part of the RP report. After the meeting I participated in email discussions to finalize the review panel summary report. This CIE report is structured according to my interpretation of the required format and content described in Annex 1 of Appendix 2.

Summary of findings

I first provide summaries that apply to both assessments, and then present stock-specific summaries where necessary.

ToR 1. Become familiar with the draft stock assessment documents, data inputs, and analytical models along with other pertinent information (e.g., previous assessments and STAR panel report when available) prior to review panel meeting.

I reviewed in detail the draft stock assessment and background documents for Dover Sole and Pacific Spiny Dogfish (including 2011 CIE Reviews) that were provided (see Appendix 1).

I also reviewed some additional documents:

Thorson, J.T., Cunningham, C.J., Jorgensen, E., Havron, A., Hulson, P.J.F., Monnahan, C.C. and von Szalay, P., 2021. The surprising sensitivity of index scale to delta-model assumptions: Recommendations for model-based index standardization. Fisheries Research, 233, p.105745.

Thorson, J.T., 2019. Guidance for decisions using the Vector Autoregressive Spatio-Temporal (VAST) package in stock, ecosystem, habitat and climate assessments. Fisheries Research, 210, pp.143-161.

ToR 2. Discuss the technical merits and deficiencies of the input data and analytical methods during the open review panel meeting.

Landings Input data

The STATs have created long time-series of landings (since 1911 for Dover Sole, and 1916 for Spiny Dogfish) which is a merit. The accuracy of estimates of landings and discards has improved over time, as expected. This is also a merit. A deficiency is that there is uncertainty in catch estimates, and more so for historic periods and when interpolations are used to fill in catches for some years. This uncertainty was not quantified and provided to the RP. There is an important need for STATs to provide information on the quality of the annual catch estimates, and more specifically to quantify the uncertainty in these estimates. As a start, this could involve a plausible range of landings and discards (i.e., bounds) that data providers agree with (e.g., DFO, 2017; DFO, 2020), such that it is considered implausible that catches could be outside the

bounds. In time-periods when the landings and discards estimates are considered to be very complete (i.e., a census) then lower and upper bounds could be the same as the estimates.

It is also not clear that the SS3 assessment modelling approach could adequately utilize information on catch uncertainty.

DFO. 2020. Stock Assessment of NAFO Subdivision 3Ps Cod. DFO Can. Sci. Advis. Sec. Sci. Advis. Rep. 2020/2018. https://waves-vagues.dfo-mpo.gc.ca/Library/40877413.pdf.

DFO. 2017. Assessment of the Atlantic Mackerel Stock for the Northwest Atlantic (Subareas 3 and 4) in 2016. DFO Can. Sci. Advis. Sec. Sci. Advis. Rep. 2017/034. https://waves-vagues.dfo-mpo.gc.ca/Library/40619576.pdf.

A. Dover Sole

The catches were treated differently in the 2021 assessment compared to the 2011 assessment. In the 2021 assessment, commercial removals for all gear types were combined into two area-specific fleets: a California fleet and a combined Oregon/Washington fleet. This was done because of difficulties associated with separating data between Oregon and Washington. I conclude this is a good rationale.

Various studies were used to infer historical discards. Table 3 in the draft assessment report (Wetzel and Berger, 2021) summarize the discard rates, which I assume are in fractions of "kept" catch. Standard deviations are included, and they are very large relative to the estimates, with CV's usually more than 100%. Section 2.1.2 in Wetzel and Berger (2021) is interesting but lacks a clear description of how the numbers were derived in the 'Total Dead' column in Table 1, and this makes it difficult for me to assess the implications of the high uncertainty in Table 3 on the uncertainty in the total catch deaths in Table 1. At face value, the high uncertainty of the discard estimates in Table 3 suggests substantial uncertainty in the total dead in Table 1. However, my impression from the RP meeting was that the STAT felt the catch data were reliable. Hence, I may be mis-interpreting the Std.Dev. in Table 3. I recommend that discard rate confidence intervals be provided for Table 3, and that these confidence intervals also be incorporated in Table 1.

B. Pacific Spiny Dogfish

The descriptions of the methods used to estimate landings and discards were extensive. A new method was briefly described to predict trawl discards based on Sable fish total catches. However, the details of this were not clear to the RP and several requests were provided by the Panel on this issue:

Request No. 1: Provide a time series plot of the residuals of the total catch relationship between sablefish and spiny dogfish from the observer data.

Request No. 4: Provide the uncertainty intervals of the spiny dogfish historical discard estimation.

Request No. 5: Provide the discard rates applied to trawl and non-trawl landings.

Request No. 6: Provide details on calculating the prediction intervals for the historical bottom trawl discards and provide the catch streams for the low and high alternative runs (from request #4).

Request No. 12: Repeat request #4 and evaluate the sensitivity of the historical discard assumptions under each catch stream when WCGBTS q is estimated. Reproduce the figures under request #4 with an accompanying table of the q values and other model outputs. Also provide the total biomass time series under each of these scenarios.

I was satisfied with the responses to these requests. However, the uncertainty intervals were implausible in that they covered negative values. These implausible negative values were truncated at zero for Requests 6 and 12. This also suggests that the upper confidence intervals are too low. In response to Request 6, the STAT indicated that they followed the same approach as was recently used for skates, and there is a publication on this which was not provided to the RP before the meeting. The STAT committed to providing similar information for Spiny Dogfish but this was not provided during the Panel meeting.

A Research Recommendation was also provided by the Panel on this issue:

Re-evaluate approaches for informing the historical discards of spiny dogfish, including examining existing literature. If the preferred method continues to be examining the total catch of spiny dogfish in association with the total catch of sablefish in recent years of at-sea observations, the sablefish catch data should be parsed to the portion of the fishery on the shelf where spiny dogfish occur by excluding trawl efforts on the slope. This could be done by excluding winter trawl effort for sablefish or by using a MacCall-Stephens approach of filtering out efforts where sablefish are caught with Dover sole and thornyheads, which is indicative of slope targeting of the DTS (Dover sole-thornyheads-sablefish) species.

I agree that this is an important recommendation. Furthermore, I recommend better documentation of analysis conducted and a better approach be investigated to provide plausible discard confidence intervals that do not cover zero.

I did not have a detailed understanding of the spatiotemporal distribution of Spiny Dogfish and the various fishing fleets that catch this species, to provide more concrete suggestions of how to improve the bycatch-discard estimation. A more thorough review of this would have taken most of the RP time; hence, this should be done as a separate 'data inputs' review process with fishery experts and by-catch estimation experts, and reviewers with expertise in this evolving research area (e.g., Benoît, 2013; Benoît et al., 2015; Bowlby et al., 2021). This is important because the SS3 assessment model fits the catch and discard data exactly, and errors in these data will propagate directly to errors in assessment model outputs.

I recommend studies be conducted on estimating discard mortality of Spiny Dogfish for both the bottom trawl and non-trawl fleets. This could include visual determinations of direct mortality, as well as studies on post-release mortality.

Bowlby, H.D., Benoît, H.P., Joyce, W., Sulikowski, J., Coelho, R., Domingo, A., Cortés, E., Hazin, F., Macias, D., Biais, G. and Santos, C., 2021. Beyond post-release mortality: inferences on recovery periods and natural mortality from electronic tagging data for discarded lamnid sharks. Frontiers in Marine Science, 8, p.325.

Benoît, H.P., Capizzano, C.W., Knotek, R.J., Rudders, D.B., Sulikowski, J.A., Dean, M.J., Hoffman, W., Zemeckis, D.R. and Mandelman, J.W., 2015. A generalized model for longitudinal short-and long-term mortality data for commercial fishery discards and recreational fishery catch-and-releases. ICES Journal of Marine Science, 72(6), pp.1834-1847.

Benoît, H.P., 2013. Two decades of annual landed and discarded catches of three southern Gulf of St Lawrence skate species estimated under multiple sources of uncertainty. ICES Journal of Marine Science, 70(3), pp.554-563.

Length Compositions

A technical merit of both the Dover Sole and Pacific Spiny Dogfish assessments is the detailed information provided on sampling for length compositions, combined with the SS3 assessment model that can use length compositions, age compositions, and length-stratified age compositions. Length compositions provide an important source of information about variation in year class strength and total mortality rates (the latter based on fleets with asymptotic selectivity).

The precision of length samples is primarily summarized using an effective (i.e., input) sample size calculation. I did not understand the basis for the calculations, but this involved both the number of fish sampled and the number of trips. In some cases (i.e., WCGOP lengths) the input sample sizes were based on the number of trips. Fishery length sampling designs are probably complex, highly stratified cluster sampling, with many strata with incomplete sampling, and the statistical properties of the composition estimates are likely difficult or impossible to derive analytically. Nonetheless, I think the uncertainty of the composition estimates, including expansions, needs to be quantified better.

Standardization of compositional data has been advocated by Thorson (2014) and related issues of "representative sampling" should be considered for Dover Sole and Spiny Dogfish. Both STAT's used VAST model-based approaches to produce improved abundance index time-series compared to more traditional design-based approaches, but the same potential problems with design-based methods exist for length- and age-compositions. A spatiotemporal modelling approach was used for Spiny Dogfish and the WCGBTS and AFSC Triennial survey sampled lengths, but this was not reviewed by the Panel.

I recommend a bootstrap re-sampling procedure (e.g., Jourdain et al., 2020) be investigated to estimate uncertainty (i.e., covariance) in survey and fishery length and age compositions. I recognize that it may be less straight-forward if there is data-borrowing for unsampled fishery

"strata" (i.e., gears, areas, seasons, etc.) compared to survey compositions. A bootstrap procedure may be conceptually straight forward and easy to implement, although computationally demanding. There may be other analytical approaches that produce the same information, and that would be acceptable as well.

Jourdain, N.O.A.S., Breivik, O., Fuglebakk, E., Aanes, S. and Vølstad, J.H., 2020. Evaluation of sampling strategies for age determination of cod (*Gadus morhua*) sampled at the North Sea International Bottom Trawl Survey. ICES Journal of Marine Science, 77(3), pp.859-869.

Thorson, J.T., 2014. Standardizing compositional data for stock assessment. ICES Journal of Marine Science, 71(5), pp.1117-1128.

A. Dover Sole

Although discard input samples sizes are sometimes low, discards seem to be fairly minor so improved sampling might not lead to a substantially improved assessment.

Retained fishery length samples seemed good overall; however, there are a few aspects that are potential technical deficiencies:

- Length sample sizes have been relatively low from CA in 2017-2020, and the majority are from one port. If landings have not followed a similar pattern, then this indicates that recent length samples may not be indicative of the catch from the CA region as a whole. This requires further investigation about between-port variation in length distributions in CA region. Also, a catch by port figure, or sampled lengths per ton of catch, similar to Fig. 11 in Wetzel and Berger (2021), should be provided.
- Few lengths samples were provided by WA in 2020.

Length information from surveys seemed good, which is a technical merit.

I also think it would be useful to investigate the utility of multi-panel "SPAY" plots (e.g., https://rpubs.com/rajeevkumar/SPAY) of length- and age-composition time-series from the various sources, to provide a pre-assessment-model summary of consistency of recruitment information among the data sources. These are just plots of standardized deviations in compositions over time and they can be useful to detect strong and weak year classes. By comparing multiple data sources, we can get a high-level understanding of the consistency of the information across the data sources.

B. Pacific Spiny Dogfish

Most catches are discarded for this species, and this is where most of the length samples come from. The trawl discard length samples seemed to have more between-year variability compared to non-trawl discards. A technical deficiency is the shortness of the length composition time-series (only since mid 2000's) and the high between-year variability of some of these data. I was uncertain about how representative the length samples were of the retained and discarded catch, which is also a technical deficiency. However, Gertseva et al. (2021) described that the "discard length composition data were expanded, to account for non-proportional sampling of spiny dogfish among hauls and trips".

A spatiotemporal modelling approach was used for Spiny Dogfish and the WCGBTS and AFSC Triennial survey sampled lengths, but this was not reviewed by the Panel. However, this is potentially a merit.

Age compositions – merits and deficiencies

Both assessments used some age information, which is a merit overall.

A. Dover Sole

There have been no fishery ages from CA since 2008. The number of otoliths read from OR+WA have been fairly low since 2009. The fishery age compositions seem uncertain, as indicated by the wide confidence intervals for mean age. This is a deficiency.

The number of fish aged in the 2019 WCGBTS survey was less than half of recent years. This resulted in a substantial decrease in the precision of the age information in 2019, as evidenced by the increase in the mean age confidence intervals in Fig. 38 in the draft assessment document (i.e., Wetzel and Berger, 2021). This is a small deficiency. Improvements in age sampling should result in an improved stock assessment.

A new ageing analysis was conducted for otoliths read by the CAP lab and CDFW. The ageing error analysis for otoliths read by the CAP lab consisted of over 8,000 double reads of Dover sole otoliths. The ageing error analysis for otoliths read by CDFW used the same data that were available in 2011. This is a merit.

B. Pacific Spiny Dogfish

A considerable amount of the RP time was spent discussing the age measurement issues for this species. There is little data available, and issues involved with using dorsal spines to estimate dogfish age is a significant source of uncertainty and bias. The lack of age data and the poor understanding of the reliability of estimated ages was identified by the RP as a technical deficiency of the assessment, which I fully agree with.

The RP identified a short-term research recommendation that needs to be investigated before a more reliable stock assessment may be realistically achieved.

The panel recommends that research be conducted to examine the in detail the aging bias issue for mature females. The Panel suggests a re-examination of existing data, models, and methods used to derive age and growth.

I fully agree with this recommendation. Furthermore, methods to extrapolate age based on worn spines need additional validation even for immature females and males.

Natural mortality rate

This is a difficult parameter to estimate based on the data typically available for stock assessment. M was assumed to be the same for all sizes and years. This seems to be a common assumption in US west coast assessments. A Prior on M was used in both assessments. The availability of this information is an assessment merit. Both the Dover Sole and Spiny Dogfish assessments fixed female M based on the median of a priors developed by Hamel (2015). The

Spiny Dogfish assessment also fixed male M, whereas the Dover Sole assessment estimated male M as an offset to female M. Additional comments about M are provided under Tor 3.

Length-weight relationship

Both draft assessment reports included figures that plotted weight versus length for individuals and included log-linear model fits. This is a merit. However, I think it is also useful to test if there is temporal or spatial variation in the weight-length relationship that could indicate variability in condition over time and/or space.

A. Dover Sole

The log-linear model fit is good, but a plot of residuals versus log-length would help me better evaluate the fit.

B. Pacific Spiny Dogfish

It was hard to tell how well the log-linear model fit the male data, but it seems there is some lack of fit for females such that the estimated model may slightly under-estimate the weight of large females.

Length-age relationship

Both models assumed time-invariant growth rates, and parameters were estimated within the assessment model. Estimating growth internally is a good way to account for the length-selectivity of the gears used to obtain the age samples, and age measurement error. Both these factors will lead to biases in VonB parameter estimates. In particular, age measurement errors will produce negative bias in Linf and positive biases in K, and the magnitude of the biases will depend on the magnitude of the measurement error (e.g., see Dey et. al., 2019). Also, if length-stratified age sampling is used to collect otoliths for aging then treating these data as conditional age distributions is a way to address the bias introduced by this sampling design for growth model estimation (e.g., see Perreault et al., 2019). However, time- and space-invariant assumptions about growth requires verification.

Perreault, A.M., Zheng, N. and Cadigan, N.G., 2019. Estimation of growth parameters based on length-stratified age samples. Canadian Journal of Fisheries and Aquatic Sciences, 77(3), pp.439-450. https://doi.org/10.1139/cjfas-2019-0129.

Rajib Dey, Noel Cadigan, and Nan Zheng. 2019. Estimation of the Von Bertanlanffy Growth Model when Ages are Measured with Error. Journal of the Royal Statistical Society, 68: 1131-1147. https://doi.org/10.1111/rssc.12340.

A. Dover Sole

Spatial and temporal variations in size at age was addressed in Wetzel and Berger (2021), but more could be done. In particular, some statistical assessment of differences in growth model parameters among areas and/or years would be useful. This might better be addressed with SS3 by examining profiles for VonB K offsets for blocks of years or spatial regions when data are disaggregated by regions.

B. Pacific Spiny Dogfish

The problems with measuring age for Spiny Dogfish need to be resolved before investigating if growth rates have changed over time or space.

Maturity and Fecundity

Similar to my growth model comments, it is important to evaluate if maturity-at-length has changed over time. Dramatic changes in age-based maturities have occurred for some east coast groundfish stocks (e.g., Zheng et al., 2020). However, the maturity-at-length relationship may be more stable than maturity-at-age.

Zheng, N., Robertson, M., Cadigan, N., Zhang, F., Morgan, J. and Wheeland, L., 2020. Spatiotemporal variation in maturation: a case study with American plaice (*Hippoglossoides platessoides*) on the Grand Bank off Newfoundland. Canadian Journal of Fisheries and Aquatic Sciences, 77(10), pp.1688-1699.

A. Dover Sole

A new coastwide estimate of functional maturity was developed for this assessment, in response to a 2011 STAR Panel research recommendation. This is a merit! Spatial differences in maturity rates were examined for the regions north and south of Point Reyes. The spatial differences seemed to be statistically significant and supported by previous studies. However, the assessment did not account for these spatial differences in maturity and assumed a homogeneous population structure for Dover sole off the West Coast due to insufficient time to adequately evaluate the viability of creating a spatial model for Dover sole. Tagging studies seem to indicate the adult Dover Sole do not move extensively, or at least have strong seasonal homing behaviors. In this case spatial variation in various aspects of stock productivity are more likely. I fully appreciate the complexity of developing a spatial stock assessment model, and I suggest that this should be considered as a longer term, research recommendation and probably as a specific group project.

B. Pacific Spiny Dogfish

The relationship between female size and maturity was taken from recently published work (Taylor and Gallucci 2009), based on 499 fish collected in Puget Sound in the 2000s. This information should periodically be collected from other areas and more recent years.

Indices of abundance

Both assessments used VAST model-based indices derived from various scientific surveys, as opposed to more traditional design-based indices. VAST model selection was based on AIC, even though Thorson et al. (2021) described that this will not always lead to the selection of the most appropriate model. However, I appreciate that Thorson et al. (2021) is a very recent paper and best practises for the routine application of VAST are still evolving.

I have two comments/criticisms of how VAST was applied for both species:

1. VAST was applied to total biomass per tow. The size structure of catches was not considered, and this likely affects local spatial variability in catch biomass. I think it is

preferable to apply VAST to catch number-at-length, with time, space, and length effects in the model.

2. VAST was applied separately to the AFSC, NWFSC slope, and WCGBTS surveys. There is a missed opportunity for a combined analysis of these surveys to create a longer index time-series. A VAST model applied to individual catch-at-length data may have more information to determine the relationship between these survey catches and local stock abundance compared to the assessment model. A research recommendation is to investigate extending the fishery independent index time-series as long as possible. Much effort has been applied to creating a long time-series of fishery catches, and I think this should also be the case for survey index time-series.

My experience with spatiotemporal models is that qqplots and KS tests are not very helpful for detecting lack of fit, even with simulated data when we know the true model. It may be unrealistic for STATs to develop a really good fitting spatiotemporal survey index standardization model for every stock. In another recent CIE review, I described some ideas on additional diagnostics that are relevant for stock assessment purposes. I have provided this in Appendix 4. My main recommendation is that STAT's should provide design-based averages of the VAST ordinary raw residuals. To some extent the STAT's did this for Dover Sole and Spiny Dogfish by comparing VAST and design-based indices of abundance. However, I was unsure if the area represented by the design-based indices was the same as the area VAST indices were integrated over. For diagnostic purposes, these areas should be the same; however, for stock assessment purposes it is potentially good that a VAST model can extrapolate to areas that are not sampled by a survey each year. Variable survey coverage is a stock assessment problem that VAST can be used to address. I recommend that STAT's provide diagnostics on 1) differences in VAST predictions versus simple design-based predictions in survey sampled areas, and 2) the total stock area decided for the assessment. The survey sampled area may differ each year, but the comparison is only for diagnostics purposes.

A. Dover Sole

A research need identified by Wetzel and Berger (2021) involved the spatiotemporal distribution patterns of length and sex ratios with depth. The VAST model aggregated across these factors.

The Q-Q plots for the WCGBTS, early Triennial, late Triennial, and AFSC slope survey diagnostics all looked problematic to me. The assessment authors simply concluded that the substantial departures from normality in the residuals (I assume they are Gaussian quantile residuals) was not meaningful. This begs the question: what is meaningful? Also, they concluded there were no clear spatial patterns in residuals, but it was not apparent from the information presented how this conclusion was reached. There are statistical tests for spatial autocorrelation in residuals that could have been applied. Nonetheless, the trends in the VAST indices were similar to the trends in the design-based indices. The VAST indices account for vessel effects, which is good, although we did not review this part of the VAST models. I am concerned that vessel and year effects could be somewhat confounded, and the correlation matrix of these effects should be examined to check for this. Surprisingly, the VAST index standard errors were often considerable larger than the design-based indices, which is the opposite outcome of the typical motivation for applying a model-based approach (e.g., Särndal et al., 2003; Breidt and

Opsomer, 2017; Skinner and Wakefield, 2017), but maybe this is because of the vessel effects which are a source of variability that is not accounted for in the design-based indices.

Breidt, F.J. and Opsomer, J.D., 2017. Model-assisted survey estimation with modern prediction techniques. Statistical Science, 32(2), pp.190-205.

Särndal, C.E., Swensson, B. and Wretman, J., 2003. Model assisted survey sampling. Springer Science & Business Media.

Skinner, C. and Wakefield, J., 2017. Introduction to the design and analysis of complex survey data. Statistical Science, 32(2), pp.165-175.

B. Pacific Spiny Dogfish

The assessment authors decided that the lognormal model was better for accounting for very large catches compared to the gamma model. They also demonstrated that assessment outputs were very similar for survey indices based on the delta-lognormal or delta-gamma models. However, there may be better alternatives to the delta-lognormal approach for this extreme catch situation (i.e., Thorson et al., 2011) that could be considered in future research. It is important to understand the mechanisms that produce the large aggregations of spiny dogfish that produce the large catches. Consider if these aggregations are likely to exist in other years but are not sampled due to chance.

Thorson, J.T., Stewart, I.J. and Punt, A.E., 2011. Accounting for fish shoals in single-and multi-species survey data using mixture distribution models. Canadian Journal of Fisheries and Aquatic Sciences, 68(9), pp.1681-1693.

ToR 3. Evaluate model assumptions, estimates, and major sources of uncertainty.

Stock Structure

A. Dover Sole

Dover sole exhibit complex seasonal and ontogenetic movement to deeper waters but also shifting seasonally, moving from shallower feeding grounds on the continental shelf during the summer months to deeper spawning habitat on the outer continental shelf and slope in the winter. There is a pattern of sex ratio by depth in the WCGBTS with more males being found in middle depths and more females found in shallow and deeper depths. However, the specific mechanisms that drive stock structure, and related variability over space and time, are not well understood. Dover sole eggs and pelagic larvae have a protracted pelagic phase. Tagging studies indicated seasonal movements but little evidence of north-south movement or appreciable mixing among PMFC statistical areas.

Similar to the previous 2011 assessment, a single coastwide population was modeled allowing for area-specific fleets (CA and OR/WA) and separate growth and mortality parameters for each sex (i.e., a two-sex model).

Stock boundaries were identified by Wetzel and Berger (2021) as a research need. Dover sole live deeper than the range of the fisheries and surveys, and this has created cryptic biomass in the assessment outcomes whose magnitude is speculative. Research into abundance in deep areas would be useful to verify that the assessment adequately predicts the entire spawning stock. The RP also recommended studies be conducted to verify the magnitude of the cryptic biomass.

Wetzel and Berger (2021) suggested that there may potentially be multiple stocks of Dover sole due to limited intermingling of adult populations, but larvae probably intermingle during their long pelagic life. Localized depletion of spawning components could be a problem. Differences in productivity (recruitment, growth, maturity, and mortality rates) among stock components may also indicate a need for spatial harvest strategies. This needs more research.

B. Pacific Spiny Dogfish

Spiny dogfish is a transboundary stock, and there are high densities of dogfish close to the U.S.-Canada border. Limiting the assessment area to the U.S. West Coast coastal waters does not allow for including a full range of spatial and temporal dynamics for the species, and therefore results may possess additional uncertainty associated with not looking at the full scope of stock's distribution. A spatial population dynamics model, which included data from several tagging studies in the Northeast Pacific Ocean, estimated movement rates of about 5% per year between the U.S. coastal sub-population of dogfish and that found along the west coast of Vancouver Island in Canada. This could accumulate to 50% exchange in 8 years.

I agree with the STAT recommendation that U.S. and Canada should explore the possibility of a joint stock assessment in future years.

Model estimation

SS3 is a flexible stock assessment modelling framework that can integrate intermittent samples of length compositions, age compositions, and various types of abundance indices. It is an appropriate modelling framework for the Dover Sole and Spiny Dogfish.

For both species, growth was estimated within the assessment model which is appropriate given the size selectivity of the fisheries and surveys. I conclude that the time-blocks of selectivity used in Dover Sole and Spiny Dogfish models was appropriate.

Model convergence was checked and was acceptable for the base models. Convergence was also checked for sensitivity models, and in a few cases problems were reported. Occasional non-convergence is expected.

I conclude from the review meeting that the model was very competently applied. The skill of the STATs with SS3 was a strong merit. I was impressed with the ability of the STATs to quickly produce relevant plots and other output based on requests for additional runs. This greatly improved the efficiency of the review.

Both models fit substantial amounts of length and age compositional data. It is common to see correlated errors in the composition residuals, such that the residuals are the same sign for neighboring lengths and age classes, and to some extend for neighboring years. SS3 assumes a Multinomial distribution for compositions and there should only be negative correlation between

residuals, although clearly there must be positive and negative residuals each year because the composition proportions must sum to one and over-estimation is balanced with under-estimation. The same issue exists for the Dirichlet-Multinomial distribution. The Multinomial and Dirichlet-Multinomial distributions are more appropriate for nominal categorical data where there is no natural ordering between categories. Therefore, the correlation between category responses is always expected to be negative for this type of data. However, length and age compositions are ordinal or even interval categorical data and we should expect different correlation patterns in this case. The impact of mis-specifying the statistical distribution of the compositions is higher negative loglikelihoods (nll's) than the data warrant, and perhaps too much weight given to fitting the compositions. SS3 uses an index likelihood function that assumes independent errors across time. If errors are autocorrelated, then the SS3 independence assumption will also result in higher nll's and perhaps too much weight given to fitting the compositions, which is a deficiency. It is reasonable to expect autocorrelated errors for fishery-dependent data because of the many possible "transitory" fishery effects. However, it is difficult to reliably estimate autocorrelation parameters and a prior on these parameters may be justified.

Binning of lengths and ages is a pragmatic approach to deal with large sampling errors and infrequent length or age categories. However, too much binning will not be good. I am not sure how flexible SS3 is in this regard. Both assessments used "regular" length bins with fixed widths when variable bin widths may have been more appropriate. Different sizes of bins, especially those that aggregate small and large lengths, may be more appropriate if they could be defined as fleet-specific.

I find it difficult to evaluate the adequacy of the fits to length compositions and conditional age compositions. I am never sure when fits are too bad to accept.

A. Dover Sole

The lengths in the population were tracked by 1 cm intervals and the length data were binned into 2 cm intervals. I suggest that binning all lengths <= 15cm or 20cm, and all lengths >=50cm, may have produced an assessment that was more robust to sampling errors at these infrequently observed sizes. A smaller plus group age could also produce a more robust assessment. However, Pearson residuals at these sizes and ages were often small so the impacts of different binning formulations may not be substantial.

B. Pacific Spiny Dogfish

The length frequency distributions were represented as thirty, 4-cm bins ranging between 12 and 132 cm. Population length bins were defined at a finer 2-cm scale, ranging between 10 and 136 cm. Possibly more robust options are to aggregate all lengths less than 20cm and greater than 110 or 120cm.

There were some large Pearson residuals for the midwater trawl fleet, the non-trawl discard fleet, and the recreational fleet. The robustness of the assessment to occasional "outliers" should be assessed. An SS3 option for robust estimation (e.g., Aeberhard et al., 2020) should be considered as a future option.

Aeberhard, WH, Cantoni, E, Field, C, Künsch, HR, Mills Flemming, J, Xu, X. Robust estimation for discrete-time state space models. Scand J Statist. 2020; 1– 21. https://doi.org/10.1111/sjos.12482

Growth

For both stocks, the size-at-age was modelled separately for the two sexes. Females and males have separate growth curves (fully estimated within the model) and sex-specific weight-at-length parameters. This seemed appropriate to me.

A. Dover Sole

The impact of time-and space-varying size-at-age needs more consideration in the assessment process.

B. Pacific Spiny Dogfish

Ageing uncertainty seems to be the dominant uncertainty in modelling size-at-age. This needs to be resolved before considering refinements to growth models.

Selectivity

Selectivity was modelled via fleets and time-blocks, which is typical in US stock assessments. The choice of time-blocks was decided based on knowledge of changes in management regulations and by residual analyses and model building, aided by AIC and assessment improvement in fit relative to model complexity. Selectivity was modeled as a function of length, using 6 parameter double-normal selectivity curves. This is good, and better than modeling selectivity as a function of age.

The efficacy and robustness of parametric selectivity models is always a concern. Double-normal domed selectivity patterns often look implausible to me. Time-blocking of selectivity is also tedious but useful when there are important changes in management measures. In other fora (e.g., Canada, ICES) this type of blocking is not commonly done. Selectivity is modelled annually but sometimes with smooth variations over time and age /length (e.g., correlated random walk). Such an option may be useful for SS3, for diagnostic purposes at least. This is a nonparametric approach that will substantially simplify formulation and review of assessment models. The SS3 approach requires much consideration of which fleets to model, what parametric selectivity model to choose, and what time blocks are appropriate. If a STAT gets this right then the assessment may be slightly more efficient and reliable compared to a random-effects approach, but if the STAT gets this wrong then the reverse may occur and the assessment may be much less reliable. However, to do this effectively will probably require a random effects modelling approach and simulation testing for data scenarios relevant to US stock assessments, so this is a longer-term suggestion.

Fleets and surveys with asymptotic selectivity over ranges of lengths that are frequently caught contribute more direct information about mortality rates.

A. Dover Sole

This stock had a substantial amount of length and age composition information available, but selection functions were complex and had some degree of dome pattern. This reduces the mortality rate information provided by the length compositions. Selectivity for some fleets was based on spline selectivity models, which seemed more plausible. I am uncertain why this approach is not a default, or used to better motivate parametric selectivity models. However, appropriate knot placement in areas of high function curvature or when data is more dense is important for more reliable spline models.

B. Pacific Spiny Dogfish

Only the non-trawl landings and recreational fleets were assumed to have asymptotic selectivity for males and females. The non-trawl landings female lengths with asymptotic selectivity also had substantial sample sizes and were probably informative about total mortality rates (Z's). The lengths that had asymptotic selection for non-trawl landings males and recreational catches (males+females) had low sample sizes and these data are probably not informative about Z. AFSC Slope and NWFSC Slope Surveys for males had asymptotic selectivity at about 60cm+ and this information should be somewhat informative about mortality rates. However, the length composition time-series for this stock is relatively short, and since most fleets and sexes have domed selectivity, the size composition information do not seem very informative about mortality rates.

Natural Mortality

I do not agree with fixing M. I think the priors should be used for their intended purpose and not to provide fixed values for M.

These priors are built from M relationships with other life-history parameters. The M estimates contributing to the priors are probably derived in many cases from less information than available for Dover Sole and Spiny Dogfish. "Borrowing" relevant information from the assessments of similar stocks is justified for parameters that an assessment is uninformative about. However, if an assessment is informative about M, then this information should not be discarded simply because the assessment estimate differs from the prior median. I recognize that in some cases a stock assessment model will produce a biased estimate of M because some other aspect of the model is mis-specified. In this case, assessment authors should focus some investigation on the source of possible misspecification. I understand that time is limited and there are many things to prepare for in a stock assessment, Hence, I recommend that improved M diagnostics be investigated in general (i.e., not specifically for Dover Sole and Spiny Dogfish) with the goal of helping STATs find the best available model specifications for their stock. There will be a CAPAM workshop in June 2021, and M diagnostics are on the agenda, so hopefully there will be some specific recommendations on M diagnostics that could be applied.

A. Dover Sole

The STAT concluded in their draft assessment documents that "the estimate of female natural mortality was deemed to be unrealistic due to low parameter estimates (around 0.08 yr^{-1}) which did not appear to be supported by the data". In fact, the assessment data was highly informative about M. The length, age, and index data all indicated M in the range of 0.07-0.09. For the length compositions, the profile for M was primarily driven by data from the AFSC slope survey. Other

length data were fairly uninformative about M. Both profiles of the age and index data from the WCBTS "favored" a low (<=0.07) value of M. There is a poor fit to the WCGBT age data (see below) which may be the reason why this data is supporting a lower value. Only the age compositions from CA and OR_WA supported the female M=0.108 that was fixed in the model. However, there is little recent age data available for CA and OR_WA. I conclude that the specific of M for Dover Sole is not well supported by the available data.

The low and high states of nature chosen for this stock correspond to female M of 0.084 and 0.126 per year. I agree with the high state but the low state may not be low enough.

B. Pacific Spiny Dogfish

Given the relatively short time-series of length data and sparse age data, it is not surprising that the assessment model provided little information about M. Signals were contradictory as well, with indices indicating a higher M and lengths compositions indicating a lower M value. However, these data sources on their own were not very informative, and less so together.

It seems that this is a species that will continue to be difficult to age. The species is primarily discarded so a sampling program to estimate the length distributions of total discards is also a challenge. Hence, I am not anticipating the information base for M will improve in the near future. There are many published studies where pop-up satellite tags have been used to estimate M for shark species, and this seems like a practical short-term project that could produce direct information about M. However, these tags are expensive.

Stock-recruitment

A. Dover Sole

The assessment assumed a Beverton-Holt stock-recruitment relationship. Steepness was fixed at 0.80, the mean of the prior. A sensitivity analysis and a likelihood profile were performed for steepness. Recruitment variability was fixed (σ_R = 0.35) based on the estimated variation in recruitment from the base model. Recruitment deviations were estimated from 1880 - 2018 to appropriately quantify uncertainty in the early model years.

Likelihood profiles for steepness were approximately flat for steepness values greater than 0.5, and the assessment model could not reliably estimate how large steepness was. The fixed value of 0.8 seemed reasonable. The fraction unfished is sensitive to the steepness value, but stock status evaluations relative to the management target were approximately the same. The treatment of recruitment deviations was sensible. Given the amount of length and age composition data used in the model, I would have expected the model to be more informative about σ_R. Since this parameter was fixed, a sensitivity analyses and likelihood profile should have been provided for this, but I have no reason to expect that the assessment outcomes would differ substantially to alternative and reasonable values for σ_R.

B. Pacific Spiny Dogfish

The spawner-recruit relationship was modeled using a functional form which allows a more explicit modeling of pre-recruit survival between the stage during which embryos can be counted in pregnant females to their recruitment as age 0 dogfish. The recruits were taken

deterministically from the stock-recruit curve since the relatively large size of dogfish pups at birth (20-30cm) suggest that variability in recruitment would be lower than for a species with a larval stage, which is subject to higher mortality rates. However, the values of the stock-recruit parameters were identified by the STAT as a major uncertainty.

The parameters controlling the relationship are equilibrium recruitment (R0), a parameter controlling the potential decrease in pre-recruit mortality as spawning output is reduced (z_{frac}), and a parameter controlling the shape of the mortality-depletion relationship (β). The base model used the survival-based relationship with fixed $z_{frac} = 0.4$ and $\beta = 1.0$. A sensitivity analysis was conducted for the values of z_{frac} and β, and assessment results were approximately the same. The two sensitivity analyses for z_{frac} and β were chosen after running models spanning a grid of values in both dimensions and choosing the combinations that produced the results most different from the base model. I like this type of structured sensitivity analysis. The lack of sensitivity is a merit of the assessment since these parameters are difficult to estimate given the available data for this stock.

The data available for this stock may not provide reliable tracking of strong and weak year classes, plus the Spiny Dogfish recruitment strategy will not produce high recruitment variability either. Hence, cohort dynamics may be obscure, and this seems like a stock that it will be inherently difficult to distinguish between mortality and selectivity effects, which is often easier to do when there is high recruitment variability and strong cohort effects in composition data.

Uncertainty

This was quantified using hessian-based standard errors and sensitivity analyses. I conclude that this was done well for both stocks. Retrospective analyses were provided and these can also give some indication of the uncertainty of key assessment outputs. Both assessments did not have large retrospective patterns.

Sensitivities

Both STATs performed a wide range of sensitivity analyses (both to data and structural model uncertainty) before the RP and documented the results in succinct and easy to understand comparison plots. However, for Dover Sole I recommend in future assessments that the STAT provide a sensitivity analyses and likelihood profile for σ_R.

ToR 4. Provide constructive suggestions for current improvements if technical deficiencies or major sources of uncertainty are identified.

A. Dover Sole

Technical deficiencies identified by the RP are:

- There were limited new fishery age data since 2010; no otoliths collected in CA after 2009 were read; limited otoliths collected in OR and WA were read.

- The number of otoliths collected and read ages from WCGBTS was reduced by about 50% in 2020 compared to previous years. This drop in available ages in 2020 is due to reduced survey effort (2 vessels versus the 4 vessels in earlier years).

My obvious recommendation is to collect and read more ages.

Unresolved problems identified by the RP are:

- The low estimate of M the model produces indicates some other model misspecification.
- There is some lack of fit to CA fishery length compositions during 2000-2020 (see Request No. 2) and WCGBTS age compositions since 2010 (see Request No. 3).

I recommend additional diagnostics analyses be investigated to determine more specifically the sources of data and years that are better fit with lower M's than the median of the prior. If no model misspecifications are discovered, then I suggest that female M should be estimated using the M prior.

The CA fishery length compositions during 2000-2020 should be further investigated to see if they are representative of the CA fleets. Alternative sources of the lack of fit could involve growth model misspecification, and I suggest spatiotemporal analyses of variability in size-at-age should be conducted to provide some insights about the potential for growth model misspecification.

Major uncertainties identified by the RP involved:

- The veracity of the cryptic biomass.
- The level of natural mortality rates.
- Stock structure and spatial productivity dynamics.

I am not a Dover Sole expert and I do not think it is useful for me to provide recommendations on how to reduce these uncertainties about cryptic biomass, stock structure and spatial dynamics. A well-designed long-term tagging program can produce direct information about M. Detailed size composition sampling from no-take zones may also yield information about M.

B. Pacific Spiny Dogfish

Technical deficiencies identified by the RP are:

- Model scale is very sensitive to assumptions on M and q.
- Ageing uncertainty and bias of older dogfish.

I do not have additional constructive suggestions beyond the research recommendations of the RP.

Unresolved problems identified by the RP are:

- Uncertainties in the aging of older, larger females.
- There is a possible error with the reference used to set M.
- Discards in all fleets between 1960 and 2002.

The possible error in M identified by the RP was based on the idea that longer-lived elasmobranchs tend to have a lower M/k ratio than indicated by the Spiny Dogfish assessment. I am not familiar with the literature on this issue.

The RP did not review the research that motivated the approach used to estimate discards; However, as I indicated above, this should be done as a separate 'data inputs' review process.

Major uncertainties identified by the RP involved:

- There are rather large uncertainties around catchability (q) from the WCGBTS.

ToR 5. Determine whether the science reviewed is considered to be the best scientific information available.

I concur with the STAR RP that the assessments for Dover Sole and Pacific Spiny Dogfish constitute the best available scientific information on the current status of the stocks and that the assessments provides a suitable basis for management decisions.

ToR 6. When possible, provide specific suggestions for future improvements in any relevant aspects of data collection and treatment, modeling approaches and technical issues, differentiating between the short-term and longer-term time frame.

I agree with the RP recommendations, which I first provide for completeness. I follow these with my additional suggestions for improvements.

Review Panel Recommendations

A. Dover Sole

Higher priority

- Consider studies to verify the magnitude of the cryptic biomass.
- Improved understanding of survey catchability could be provided via trawl escapement and herding studies. This is linked to a 2011 recommendation.
- Improved size and age fishery sampling south of Pt. Reyes should be provided, to investigate possible differences in age, size, and sex structure by depth and latitude. More generally, increase collection and reading of age compositions for the fishery to improve the application of an age structured assessment model.
- Investigate the spatial and temporal dynamics, seasonality, and ontogenetic movement that could help to capture what is happening with Dover Sole regarding the distribution of ages in the bottom trawl survey. Investigate if there is seasonality or annual environmental factors that could potentially change distribution patterns and how those pattern changes overlap with the bottom trawl survey.

Lower priority

- Consider using the AFSC Slope Survey age data as conditional age-at-lengths.
- Conduct spatiotemporal analysis of maturity-at-length and length-at-age, and examine if trends are significantly different. This is linked to a 2011 recommendation.
- Conduct additional genetic and tagging studies to examine stock structure and connectivity of the stock across its whole range.
- Consider if existing tagging information provides useful assessment information about growth and/or mortality rates.

B. Pacific Spiny Dogfish

Research to be done prior to the next assessment attempt

- The Panel also supported the STAT's recommendation that all ongoing data streams used in this assessment be continued or increased including fishery dependent sampling for length, age, and maturity, as well as fishery independent collection and aging. Fishery dependent samples should be collected in light of changing fleet dynamics and to fully cover the range of the current fishery.
- Re-evaluate approaches for informing the historical discards of spiny dogfish, including examining existing literature. If the preferred method continues to be examining the total catch of spiny dogfish in association with the total catch of sablefish in recent years of at-sea observations, the sablefish catch data should be parsed to the portion of the fishery on the shelf where spiny dogfish occur by excluding trawl efforts on the slope. This could be done by excluding winter trawl effort for sablefish or by using a MacCall-Stephens approach of filtering out efforts where sablefish are caught with Dover sole and thornyheads, which is indicative of slope targeting of the DTS (Dover sole-thornyheads-sablefish) species.
- As also recommended by the STAT, the Panel suggests that a vigorous examination of natural mortality via meta-analysis be conducted to help in establishing informative priors for M for future assessments. This analysis should be linked to other parameters such as growth.
- Like most other assessments, estimates of catchability (q) is a major source of uncertainty and an unresolved issue for this assessment. This is especially true for dogfish as they appear to be semi-pelagic and may not be available to the survey trawl consistently. As such, both the STAT and the Panel recommend future research into the catchability of dogfish in the WCGBTS. These could include depletion studies, video surveillance of trawl operations, or other analysis as appropriate bench-top analysis of co-occurring fishery dependent/independent data.
- Given the issue that worn spines of older females may produce an aging bias, the panel recommends that research be conducted to examine this issue in detail. The Panel suggests a re-examination of existing data, models, and methods used to derive age and growth.

Research needed at some point in the future.

- Given the densities of large schools of dogfish adjacent to the US - Canada border. The Panel supported the STAT recommendation that the next assessment be conducted jointly with the Canadian DFO as a potentially transboundary assessment. Prior to that, research on tagging might be helpful in either reaffirming the current 5% straying rate, or updating it.

- As outlined in the assessment report, efforts should be devoted to both improving current ageing techniques based on dogfish spines and developing new methods using other age structures. Ideally, an alternative method of ageing dogfish that does not rely on the estimation of ages missing from worn spines may be necessary. Improvement in ageing would contribute to better understanding of spiny dogfish longevity and help estimating natural mortality as well as inform growth parameters within the assessment model.

My Additional Research Recommendations

Short-term

1. There is uncertainty in catch estimates, and more so for historic periods and when interpolations are used to fill in catches for some years. There is an important need for STATs to provide information on the quality of the annual catch estimates, and more specifically to quantify the uncertainty in these estimates.
2. Investigate the utility of multi-panel "SPAY" plots (e.g., https://rpubs.com/rajeevkumar/SPAY) of length- and age-composition time-series from the various sources, to provide a pre-assessment-model summary of consistency of recruitment information among the data sources.
3. Time- and space-invariant assumptions about growth rates in the SS3 models for both stocks requires verification. I am not very familiar with the elasmobranch assessment literature on this topic, but this assumption for a flatfish stock is unusual in my experience.
4. A similar recommendation about maturity-at-length applies.
5. STAT's should provide design-based averages of the VAST ordinary raw residuals.
6. STAT's should provide diagnostics on differences in VAST predictions versus simple design-based predictions in survey sampled areas and also the total stock area decided for the assessment.
7. When vessel effects are included in a VAST index standardization model, vessel and year effects could be somewhat confounded, and the correlation matrix of these effects should be examined to check for this.
8. Priors on M should be used for their intended purpose and not to provide fixed values for M.

Long-term

1. Standardization of compositional data has been advocated by Thorson (2014) and related issues of "representative sampling" should be considered for Dover Sole and Spiny Dogfish.
2. A bootstrap re-sampling procedure or some similar procedure should be investigated to estimate uncertainty (i.e., covariance) in survey and fishery length and age compositions. The covariance will not be like the expectations from the Multinomial or the Dirichlet-multinomial distributions, but the re-sampling-based covariance may give some indication of more appropriate statistical distributions for the composition data.

3. VAST was applied to total biomass per tow. The size structure of catches was not taken into account, and this likely affects local spatial variability in catch biomass. I recommend investigations of applying VAST to catch number at length, with time, space, and lengths effects in the model. This model could be used to derive survey indices (total abundance) and length/age compositions.
4. VAST was applied separately to the AFSC, NWFSC slope, and WCGBTS surveys. There is a missed opportunity for a combined analysis of these surveys to create a longer index time-series. I recommend an investigation of extending the fishery independent index time-series as long as possible.
5. The efficacy and robustness of parametric selectivity models is always a concern. Double-normal domed selectivity patterns often look implausible to me. Time-blocking of selectivity is also tedious but useful when there are important changes in management measures. In other fora (e.g., Canada, ICES) this type of blocking is not commonly done. Selectivity is modelled annually but sometimes with smooth variations over time and age/length (e.g., correlated random walk). Such an option may be useful for SS3, for diagnostic purposes at least. This is a nonparametric approach that will substantially simplify formulation and review of assessment models. The SS3 approach requires much consideration of which fleets to model, what parametric selectivity model to choose, and what time blocks are appropriate. However, to do this effectively will probably require a random effects modelling approach and simulation testing for data scenarios relevant to US stock assessments.
6. Improved M diagnostics should be investigated in general (i.e., not specifically for Dover Sole and Spiny Dogfish) with the goal of helping STATs find the best available model specifications for their stock.

A. Dover Sole

<u>Short-term</u>

1. Discard rate confidence intervals be provided for Table 3 of the draft assessment document, and these confidence intervals also be incorporated in Table 1.
2. Length sample sizes have been relatively low from CA in 2017-2020, and the majority are from one port. If landings have not followed a similar pattern, then this indicates that recent length samples may not be indicative of the catch from the CA region as a whole. This requires further investigation about between-port variation in length distributions in CA region. Also, a catch by port figure, or sampled lengths per ton of catch, similar to Fig. 11 in Wetzel and Berger (2021) should be provided.
3. σ_R was fixed in the assessment model, so a sensitivity analysis and likelihood profile should have been provided for this.

<u>Long-term</u>

1. develop a spatial stock assessment model, including spatiotemporal variability in all relevant productivity processes (birth, growth, maturation, mortality).

2. Localized depletion of spawning components could be a problem. Differences in productivity among stock components may also indicate a need for spatial harvest strategies. This needs more research.

B. Pacific Spiny Dogfish

Short-term

1. Better documentation of analyses conducted to estimate discards, and
2. a better approach be investigated to provide plausible discard confidence intervals that do not cover zero.
3. There is some lack of fit in the log-linear weight-length model for female such that the estimated model may slightly under-estimate the weight of large females. This should be improved.
4. I do not anticipate the information base for M will improve in the near future. There are many published studies where pop-up satellite tags have been used to estimate M for shark species, and this seems like a practical short-term project that could produce direct information about M.

Long-term

1. I agree with the STAT recommendation that the U.S. and Canada should explore the possibility of a joint stock assessment in future years.
2. Conduct studies to estimate discard mortality of Spiny Dogfish for both the bottom trawl and non-trawl fleets. This could include visual determinations of direct mortality, as well as studies on post-release mortality.
3. The relationship between female size and maturity was taken from recently published work (Taylor and Gallucci 2009), based on 499 fish collected in Puget Sound in the 2000s. This information should periodically be collected from other areas and more recent years.
4. There may be better alternatives to the delta-lognormal approach for the extreme catch situation (i.e., Thorson et al., 2011) that could be considered in future research.
5. The robustness of the assessment to occasional "outliers" should be assessed. An SS3 option for robust estimation (e.g., Aeberhard et al., 2020) should be considered as a future option.

ToR 7. Provide a brief description on panel review proceedings highlighting pertinent discussions, issues, effectiveness, and recommendations.

The STAT teams provided well-structured presentations of the assessment and the very competent work completed before and during the STAR meeting. I felt that the RP was effective.

A. Dover Sole

Key discussions involved:

- time block for CA selectivity.
- poor fit at the end of the WCGBTS time series.

- likelihood profile of M including the priors.
- decision tables with the low and high states.
- technical merits and deficiencies.
- Unresolved Problems:
    - The low estimate of M the model produces indicates some other model misspecification.
    - lack of fit to CA fishery length compositions during 2000-2020 and WCGBTS age compositions since 2010.
- Major Uncertainties:
    - ontogenetic changes in the spatial distribution of dover sole that are different for males and females.
    - Uncertainty about the level of M.
    - Stock structure and spatial productivity dynamics are not well understood.
- High priority recommendations for future research and data collection:
    - studies to verify the magnitude of the cryptic biomass.
    - Improved understanding of survey catchability.
    - Improved size and age fishery sampling south of Pt. Reyes.
    - Investigate the spatial and temporal dynamics, seasonality, and ontogenetic movement that could help to capture what is happening with Dover Sole and the distribution of ages in the bottom trawl survey.

B. Pacific Spiny Dogfish

Key discussions involved:

- total catch relationship between sablefish and spiny dogfish from the observer data.
- the 80 cm cutoff in the growth function.
- uncertainty intervals of the spiny dogfish historical discard estimation.
- discard rates applied to trawl and non-trawl landings.
- sensitivity to the estimated female VonB k.
- runs where female M is estimated and WCGBTS q is estimated and fixed.
- runs where ln(R0) is the axis of uncertainty with WCGBTS q estimated with an accompanying likelihood profile.
- evaluate the sensitivity of the historical discard assumptions under each catch stream when WCGBTS q is estimated.
- decision tables and alternative models to bracket uncertainty.
- technical merits and deficiencies.
- Unresolved Problems:
    - uncertainties in the aging of older, larger females.
- Major Uncertainties:
    - catchability (q) of the WCGBTS.
    - Uncertainty about the level of natural mortality.
    - discards in all fleets between 1960 and 2002.
- High priority recommendations for future research and data collection:
    - Improved size composition sampling.
    - Re-evaluate approaches for informing the historical discards of spiny dogfish.

- a vigorous examination of natural mortality via meta-analysis.
- catchability of dogfish in the WCGBTS.
- ageing research.

Conclusions and Recommendations

Recommendations are provided under ToR 6.

However, an additional process recommendation involves presentation of results. An issue for me is that the draft assessment documents are large and it remains a struggle to locate results. There are no easy solutions to this problem; however, in a few other reviews I have participated in the r4ss outputs were provided to the RP in folders that I could navigate through a little more quickly than the assessment documents. This would include the base model and some sensitivity runs, although it may be impractical to provide this output for a large number of runs. It would be even better if assessment outputs were arranged in a reasonably small number of subfolders (i.e., stock results, length residuals, age residuals, etc.). There are a variety of document navigation tools that could also facilitate finding results.

ToR 1. Become familiar with the draft stock assessment documents, data inputs, and analytical models along with other pertinent information (e.g. previous assessments and STAR panel report when available) prior to review panel meeting.

I reviewed in detail the draft stock assessment and background documents for Dover Sole and Pacific Spiny Dogfish (including 2011 CIE Reviews) that were provided (see Appendix 1).

ToR 2. Discuss the technical merits and deficiencies of the input data and analytical methods during the open review panel meeting.

Technical merits

- STATs have created long time-series of landings (since 1911 for Dover Sole, and 1916 for Spiny Dogfish).
- The accuracy of estimates of landings and discards has improved over time, as expected.
- Both the Dover Sole and Pacific Spiny Dogfish assessments provided detailed information on sampling for length compositions.
- SS3 is a flexible stock assessment modelling framework that can integrate intermittent samples of length compositions, age compositions, and various types of abundance indices. This model was very competently applied for both stocks.
- Dover Sole retained fishery lengths samples seemed good overall.
- Both assessments used some age information.
- A new ageing analysis was conducted for Dover Sole.
- Both assessments utilized some information from M priors.
- Both assessments included figures that plotted weight versus length for individuals and included log-linear model fits.

- A new coastwide estimate of functional maturity was developed for the Dover Sole assessment.
- VAST survey index standardizations models accounted for vessel effects.
- Both assessments did not have large retrospective patterns.

Technical deficiencies

- There is uncertainty in catch estimates, and more so for historic periods and when interpolations are used to fill in catches for some years. This uncertainty was not quantified and provided to the RP.
- Dover Sole length sample sizes have been relatively low from CA in 2017-2020, and the majority are from one port.
- Spiny Dogfish length composition time-series is relatively short (only since mid 2000's) and there is high between-year variability in some of these data.
- There have been no Dover Sole fishery ages from CA since 2008. The number of otoliths read from OR+WA have been fairly low since 2009. The fishery age compositions seem uncertain, as indicated by the wide confidence intervals for mean age.
- Not enough Spiny Dogfish age data and a poor understanding of the reliability of estimated ages.
- Both assessments fixed female M, and did not incorporate the M prior for model inferences.
- Both assessments assumed space- and time-invariant growth rates, which is an important assumption that requires verification.
- The Dover Sole assessment includes substantial cryptic biomass which is difficult to verify.

ToR 3. Evaluate model assumptions, estimates, and major sources of uncertainty.

- SS3 is an appropriate assessment package for these stocks.
- For both species, growth was estimated within the assessment model which is appropriate given the size selectivity of the fisheries and surveys.
- The time-blocks of selectivity used in the Dover Sole and Spiny Dogfish models was appropriate.
- For both stocks, the size-at-age was modelled separately for the two sexes. Females and males have separate growth curves (fully estimated within the model) and sex-specific weight-at-length parameters, which seemed appropriate.
- Uncertainty was quantified well for both stocks.

Dover Sole

- Research into abundance in deep areas would be useful to verify that the assessment adequately predicts the entire spawning stock of Dover sole. The RP also recommended studies be conducted to verify the magnitude of the cryptic biomass.
- Localized depletion of spawning components could be a problem. Differences in productivity (recruitment, growth, maturity, and mortality rates) among stock components may also indicate a need for spatial harvest strategies.
- The impact of time-and space-varying size-at-age needs more consideration in the assessment process.
- The choice of female M was not well supported by the available data.

Pacific Spiny Dogfish

- The U.S. and Canada should explore the possibility of a joint stock assessment in future years.
- The robustness of the assessment to occasional length composition "outliers" should be assessed.
- Ageing uncertainty seems to be the dominant uncertainty in modelling size-at-age. This needs to be resolved before considering refinements to growth models.

ToR 4. Provide constructive suggestions for current improvements if technical deficiencies or major sources of uncertainty are identified.

Dover Sole

- Collect and read more ages.
- Additional diagnostics analyses should be investigated to determine more specifically the sources of data and years that are better fit with lower M's than the median of the prior. If no model misspecifications are discovered, then I suggest that female M should be estimated using the M prior.
- The CA fishery length compositions during 2000-2020 should be further investigated to see if they are representative of the CA fleets.
- Growth model misspecification is a possible source of the lack of fit to CA fishery length compositions during 2000-2020 and WCGBTS age compositions since 2010.

Pacific Spiny Dogfish

I do not have additional constructive suggestions beyond the research recommendations of the RP.

ToR 5. Determine whether the science reviewed is considered to be the best scientific information available.

I concur with the STAR RP that the assessments for Dover Sole and Pacific Spiny Dogfish constitute the best available scientific information on the current status of the stock(s) and that the assessments provides a suitable basis for management decisions.

ToR 6. When possible, provide specific suggestions for future improvements in any relevant aspects of data collection and treatment, modeling approaches and technical issues, differentiating between the short-term and longer-term time frame.

Summarized above.

ToR 7. Provide a brief description on panel review proceedings highlighting pertinent discussions, issues, effectiveness, and recommendations.

Summarized above.

Appendix 1: Bibliography of materials provided for review

Draft and Background Documents
Stock Assessment Review (STAR) Panel 1. Dover Sole and Pacific Spiny Dogfish

Meeting Materials:

Proposed Agenda. Stock Assessment Review (STAR) of Dover Sole and Spiny Dogfish. Pacific Fishery Management Council.

Draft Stock Assessment Documents

Vladlena Gertseva, Ian Taylor, John Wallace, and Sean E. Matson. 2021. Status of the spiny dogfish shark resource off the continental U.S. Pacific Coast in 2021. Pacific Fishery Management Council, Portland, OR.

Chantel R. Wetzel, and Aaron M. Berger. 2021. Status of Dover sole (*Microstomus pacificus*) along the U.S. West Coast in 2021. Pacific Fishery Management Council, Portland, OR.

Background Materials

Acronyms Used in West Coast Groundfish Assessments. Pacific Fishery Management Council, Portland, OR.

Methot, R.D., Wetzel, C.R., 2013. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Fisheries Research 142, 86–99.

Stacey Miller, Andi Stephens, Curt Whitmire, and Jim Hastie. 2021. Overview of West Coast Groundfish Fishery-Independent Surveys.

Richard D. Methot Jr., Chantel R. Wetzel, Ian G. Taylor, and Kathryn Doering. 2020. Stock Synthesis User Manual Version 3.30.16.

Terms of Reference for the Groundfish and Coastal Pelagic Species Stock Assessment and Review Process for 2021-2022. Pacific Fishery Management Council. December, 2020.

Thorson, J.T. and Barnett, L.A., 2017. Comparing estimates of abundance trends and distribution shifts using single-and multispecies models of fishes and biogenic habitat. ICES Journal of Marine Science, 74(5), pp.1311-1321.

Taylor, I.G., Gertseva, V. and Matson, S.E., 2013. Spine-based ageing methods in the spiny dogfish shark, Squalus suckleyi: How they measure up. Fisheries research, 147, pp.83-92.

Taylor, I.G., Gertseva, V., Methot Jr, R.D. and Maunder, M.N., 2013. A stock–recruitment relationship based on pre-recruit survival, illustrated with application to spiny dogfish shark. Fisheries Research, 142, pp.15-21.

Report of the Pre-Assessment Workshop for the 2021 Stock Assessment of Spiny Dogfish. Pacific Fishery Management Council, Portland, OR.

Report of the Pre-Assessment Workshop for 2021 Groundfish Stock Assessments of Dover Sole, Copper Rockfish, Quillback Rockfish, and Squarespot Rockfish. Pacific Fishery Management Council, Portland, OR.

Spiny Dogfish STAR Panel Report. July 11-15, 2011. Pacific Fishery Management Council, Portland, OR.

Gertseva, V., Taylor, I.G. 2012. Status of the spiny dogfish shark resource off the continental U.S. Pacific Coast in 2011. Pacific Fishery Management Council, Portland, OR.

Dover Sole STAR Panel Report. July 25-29, 2011. Pacific Fishery Management Council, Portland, OR.

Hicks, Allan C., and Chantel R. Wetzel. 2011. The Status of Dover Sole (Microstomus Pacificus) Along the U.S. West Coast in 2011. Pacific Fishery Management Council, Portland, OR.

Appendix 2: CIE Statement of Work

Performance Work Statement

External Independent Peer Review by the Center for Independent Experts

Stock Assessment Review (STAR) Panel 1 - Virtual

Dover Sole and Pacific Spiny Dogfish

May 3-7, 2021

Background

The National Marine Fisheries Service (NMFS) is mandated by the Magnuson-Stevens Fishery Conservation and Management Act, Endangered Species Act, and Marine Mammal Protection Act to conserve, protect, and manage our nation's marine living resources based upon the best scientific information available (BSIA). NMFS science products, including scientific advice, are often controversial and may require timely scientific peer reviews that are strictly independent of all outside influences. A formal external process for independent expert reviews of the agency's scientific products and programs ensures their credibility. Therefore, external scientific peer reviews have been and continue to be essential to strengthening scientific quality assurance for fishery conservation and management actions.

Scientific peer review is defined as the organized review process where one or more qualified experts review scientific information to ensure quality and credibility. These expert(s) must conduct their peer review impartially, objectively, and without conflicts of interest. Each reviewer must also be independent from the development of the science, without influence from any position that the agency or constituent groups may have. Furthermore, the Office of Management and Budget (OMB), authorized by the Information Quality Act, requires all federal agencies to conduct peer reviews of highly influential and controversial science before dissemination, and that peer reviewers must be deemed qualified based on the OMB Peer Review Bulletin standards. (http://www.cio.noaa.gov/services_programs/pdfs/OMB_Peer_Review_Bulletin_m05-03.pdf).

Further information on the CIE program may be obtained from www.ciereviews.org.

Scope:

The National Marine Fisheries Service and the Pacific Fishery Management Council will hold three stock assessment review (STAR) panels and potentially one mop-up panel if needed, to evaluate and review benchmark assessments of Pacific coast groundfish stocks. The goals and objectives of the groundfish STAR process are to:

1) ensure that stock assessments represent the best available scientific information and facilitate the use of this information by the Council to adopt OFLs, ABCs, ACLs, (HGs), and ACTs;
2) meet the mandates of the Magnuson-Stevens Fisheries Conservation and Management Act (MSA) and other legal requirements;
3) follow a detailed calendar and fulfill explicit responsibilities for all participants to produce required reports and outcomes;
4) provide an independent external review of stock assessments;
5) increase understanding and acceptance of stock assessments and peer reviews by all members of the Council family;
6) identify research needed to improve assessments, reviews, and fishery management in the future; and
7) use assessment and review resources effectively and efficiently.

Benchmark stock assessments will be conducted and reviewed for the Dover sole and Pacific spiny dogfish. These stocks were identified within the top five rankings for assessment consideration during the Pacific coast groundfish regional stock assessment prioritization process, which was based on the national stock assessment prioritization framework (http://www.st.nmfs.noaa.gov/Assets/stock/documents/PrioritizingFishStockAssessments_FinalWeb.pdf).

Dover sole was last assessed in 2011, and estimated stock depletion in that year was 83.7 percent of its unfished biomass at the start of 2011 (Hicks and Wetzel, 2012). A catch-only projection update of that assessment was conducted in 2019, and estimated depletion at that time was 77.6 percent. Dover sole range from Baja California to the Bering Sea and eastern Aleutian Islands, however the assessment addresses that portion of the stock caught in the fisheries off California, Oregon, and Washington. Dover sole are highly important to the commercial fishery, however modeling difficulties arise from the fact that females grow to be much larger than males and display ontogenetic movement to deeper waters as they age, making the older females unavailable to the fishery and to the West Coast Bottom Trawl Survey. The attainment for Dover sole is constrained by the fishery for Sablefish, as well as by market considerations.

Pacific Spiny Dogfish off the U.S. West Coast was last assessed in 2011 (Gertseva and Taylor 2012), which estimated stock depletion to be 63.2 percent of unfished spawning biomass at the start of 2011. The species range is from Baja California to the Bering Sea, however the assessment addresses the portion of the stock caught in the fisheries off California, Oregon, and Washington. Seasonal movement of some dogfish between the higher-density areas off Washington and British Columbia is likely in many/most years. Because dogfish lack otoliths, traditional methods of aging used for other groundfish species are not available. Instead, dorsal spines are used to determine age. Although these spines exhibit readable annuli, they are subject to wear, over time, which increases aging uncertainty, particularly for older fish. Additionally, preparing the spines for aging is a time-consuming, multi-step process, which has severely limited the availability of age data for use in assessments. Consequently, age data were not included

directly in the 2011 model, nor will that be the case in the new assessment. Pacific Spiny Dogfish are sporadically targeted, but are more often a bycatch species in the commercial trawl fishery, with the vast majority of retained catch being exported, mostly to Asian markets.

Assessments for these stocks will provide the basis for the management of the groundfish fisheries off the West Coast of the U.S., including providing scientific basis for setting OFLs and ABCs as mandated by the Magnuson-Stevens Act. The technical review will take place during a formal, public, multiple-day virtual meeting of fishery stock assessment experts. Participation of an external, independent reviewer is an essential part of the review process. The Terms of Reference (ToRs) of the peer review are attached in **Annex 2**. The tentative agenda of the panel review meeting is attached in **Annex 3**.

Requirements:

Two CIE reviewers will participate in the stock assessment review panel. One CIE reviewer, requested herein, shall conduct an impartial and independent peer review of the assessments described above and in accordance with the Performance Work Statement (PWS) and ToRs herein. Additionally, one "common" CIE reviewer will participate in all STAR panels held in 2021 and the PWS and ToRs for the "common" CIE reviewer are included in **Attachment A**.

The CIE reviewers shall be active and engaged participants throughout panel discussions and able to voice concerns, suggestions, and improvements, while respectfully interacting with other review panel members, advisors, and stock assessment technical teams. The CIE reviewers shall have excellent communication skills in addition to working knowledge and recent experience in fish population dynamics; with experience in the integrated-analysis modeling approach, using age- and size- (and possibly spatially-) structured models, and methods for quantifying uncertainty. Familiarity with environmental, ecosystem and climatic effects on population dynamics and distribution may also be beneficial. The CIE reviewer's duties shall not exceed a maximum of 14 days to complete all work tasks of the peer review described herein.

Tasks for Reviewers:

The CIE reviewer shall complete the following tasks in accordance with the PWS and Schedule of Milestones and Deliverables herein.

Prior to the Peer Review: Upon completion of the CIE reviewer selection by the CIE Steering Committee, the CIE shall provide the CIE reviewer information (full name, title, affiliation, country, address, email) to the NMFS Contracting Officer Representative (COR), who forwards this information to the NMFS Project Contact no later than the date specified in the Schedule of Milestones and Deliverables. The CIE is responsible for providing the PWS and ToRs to the CIE reviewer. The NMFS Project Contact is responsible for providing the CIE reviewer with the background documents, reports, and other information concerning pertinent meeting arrangements. The NMFS Project Contact is also responsible for providing the Chair a copy of the PWS in advance of the panel review meeting. Any changes to the PWS or ToRs must be made through the COR prior to the commencement of the peer review.

Pre-review Background Documents: Two weeks before the peer review, the NMFS Project Contact will send (by electronic mail or make available at an FTP site) to the CIE reviewers the necessary background information and reports for the peer review. In the case where the documents need to be mailed, the NMFS Project Contact will consult with the CIE Lead Coordinator on where to send documents. CIE reviewers are responsible only for the pre-review documents that are delivered to the reviewer in accordance with the PWS scheduled deadlines specified herein. The CIE reviewer shall read all documents in preparation for the peer review.

Documents to be provided to the CIE reviewers prior to the STAR Panel meeting include:

- The current draft stock assessment reports;
- Previous stock assessments and STAR Panel reports for the assessments to be reviewed;
- The Pacific Fishery Management Council's Scientific and Statistical Committee's Terms of Reference for Stock Assessments and STAR Panel Reviews;
- Stock Synthesis (SS) Documentation;
- Additional supporting documents as available;
- An electronic copy of the data, the parameters, and the model used for the assessments (if requested by reviewer).

Test: Additionally, two weeks prior to the peer review, the CIE reviewers will participate in a test to confirm that they have the necessary technical specifications provided in advance of the panel review meeting.

Panel Review Meeting: The CIE reviewer shall conduct the independent peer review in accordance with the PWS and ToRs, and shall not serve in any other role unless specified herein. **Modifications to the PWS and ToRs cannot be made during the peer review, and any PWS or ToRs modifications prior to the peer review shall be approved by the COR and CIE Lead Coordinator.** Each CIE reviewer shall actively participate in a professional and respectful manner as a member of the review panel's virtual meeting, and their peer review tasks shall be focused on the ToRs as specified herein. The NMFS Project Contact is responsible for any facility arrangements (e.g., video or teleconference arrangements). The NMFS Project Contact is responsible for ensuring that the Chair understands the contractual role of the CIE reviewers as specified herein. The CIE Lead Coordinator can contact the Project Contact to confirm any peer review arrangements, including the meeting facility arrangements.

Contract Deliverables - Independent CIE Peer Review Reports: The CIE reviewer shall complete an independent peer review report in accordance with the PWS. Each CIE reviewer shall complete the independent peer review according to required format and content as described in **Annex 1**. The CIE reviewer shall complete the independent peer review addressing each ToR as described in **Annex 2**.

Other Tasks – Contribution to Summary Report: The CIE reviewer should assist the Chair of the panel review meeting with contributions to the Summary Report, based on the terms of reference of the review. The Chair is not provided by the CIE under this contract. A CIE reviewer is not required to reach a consensus with other members of the Panel, and should provide a brief summary of the reviewer's

views on the summary of findings and conclusions reached by the review panel in accordance with the ToRs.

Place of Performance:

The CIE reviewers shall conduct an independent peer review during the panel review meeting scheduled for the dates of May 3-7, 2021. Due to current uncertainties in the state of the COVID-19 pandemic at that time, this meeting will be conducted as a virtual meeting, with technical assistance provided by staff from the Pacific Fishery Management Council.

Period of Performance:

The period of performance shall be from the time of award through **July 2021**. The CIE reviewers' duties shall not exceed 14 days to complete all required tasks.

Schedule of Milestones and Deliverables:

CIE shall complete the tasks and deliverables described in this PWS in accordance with the following schedule.

Schedule	**Milestones and Deliverables**
Within two weeks of the award	Contractor selects and confirms reviewers. This information is sent to the COR, who then transmits this to the NMFS Project Contact
Approximately two weeks later	Contractor provides the pre-review documents to the CIE reviewers
May 3-7, 2021	**Virtual Panel Review Meeting**
Approximately two weeks later	Contractor receives draft reports
Within two weeks of receiving draft reports	Contractor submits final CIE independent peer review reports to the COR

Applicable Performance Standards

The acceptance of the contract deliverables shall be based on three performance standards:

(1) The reports shall be completed in accordance with the required formatting and content; (2) The reports shall address each TOR as specified; and (3) The reports shall be delivered as specified in the schedule of milestones and deliverables.

Travel:

No travel is necessary, as this meeting is being held remotely.

Restricted or Limited Use of Data:

The contractors may be required to sign and adhere to a non-disclosure agreement.

NMFS Project Contact:

Andi Stephens, NMFS Project Contact
National Marine Fisheries Service,
Newport, OR 97365
Andi.Stephens@noaa.gov
Phone: 843-709-9094

Annex 1: Format and Contents of CIE Independent Peer Review Report

1. The CIE independent report shall be prefaced with an Executive Summary providing a concise summary of the findings and recommendations, and specify whether the science reviewed is the best scientific information available.

2. The main body of the reviewer report shall consist of a Background, Description of the Individual Reviewer's Role in the Review Activities, Summary of Findings for each ToR in which the weaknesses and strengths are described, and Conclusions and Recommendations in accordance with the ToRs.

   a. Reviewers should describe in their own words the review activities completed during the panel review meeting, including providing a brief summary of findings, of the science, conclusions, and recommendations.

   b. Reviewers should discuss their independent views on each ToR even if these were consistent with those of other panelists, and especially where there were divergent views.

   c. Reviewers should elaborate on any points raised in the Summary Report that they feel might require further clarification.

   d. Reviewers shall provide a critique of the NMFS review process, including suggestions for improvements of both process and products.

   e. The CIE independent report shall be a stand-alone document for others to understand the weaknesses and strengths of the science reviewed, regardless of whether or not they read the summary report. The CIE independent report shall be an independent peer review of each ToRs, and shall not simply repeat the contents of the summary report.

3. The reviewer report shall include the following appendices:

   Appendix 1: Bibliography of materials provided for review
   Appendix 2: A copy of the CIE Performance Work Statement
   Appendix 3: Panel Membership or other pertinent information from the panel review meeting.

Annex 2: Terms of Reference for the Peer Review

Stock Assessment Review (STAR) Panel 1

The specific responsibilities of the STAR panel are to:

1. Become familiar with the draft stock assessment documents, data inputs, and analytical models along with other pertinent information (e.g., previous assessments and STAR panel report when available) prior to review panel meeting.
2. Discuss the technical merits and deficiencies of the input data and analytical methods during the open review panel meeting.
3. Evaluate model assumptions, estimates, and major sources of uncertainty.
4. Provide constructive suggestions for current improvements if technical deficiencies or major sources of uncertainty are identified.
5. Determine whether the science reviewed is considered to be the best scientific information available.
6. When possible, provide specific suggestions for future improvements in any relevant aspects of data collection and treatment, modeling approaches and technical issues, differentiating between the short-term and longer-term time frame.
7. Provide a brief description on panel review proceedings highlighting pertinent discussions, issues, effectiveness, and recommendations.

Annex 3: Agenda

PROPOSED AGENDA

Stock Assessment Review (STAR) of Dover Sole and Spiny Dogfish

Pacific Fishery Management
Council Via Webinar

All Times are Pacific Daylight Time and Subject to Change During the Course of the Meeting
at the Discretion of the STAR Panel Chair

May 3-7, 2021

Monday, May 3, 2021 – 8:30 AM

Early Log-In to Resolve Connection Issues

(8:30 a.m.)

Welcome and Introductions

1. Roll Call and Introductions — Theresa Tsou, Chair
2. Review Terms of Reference — Theresa Tsou
3. Review and Approve Agenda
4. Review Virtual Format Operational Guidelines — John DeVore
5. Assign Writing Duties — Theresa Tsou

(8:45 a.m.)

Overview of the Spiny Dogfish Assessment

(9:30 a.m.)

1. Biology, Fisheries, Data, and Inputs Used — Vlada Gertseva

BREAK (10:00 – 10:15 a.m.)

2. Assessment Modeling, Performance, and Current Status — Vlada Gertseva
3. STAR Panel Requests to the Stock Assessment Team

(STAT-1) LUNCH BREAK (12:30 – 1:30 p.m.)

Overview of the Dover Sole Assessment

(1:30 p.m.)

1. Biology, Fisheries, Data, and Inputs Used — Chantel Wetzel & Aaron Berger
2. Assessment Modeling, Performance, and Current Status — Chantel Wetzel & Aaron Berger

BREAK (3:00 – 3:15 p.m.)

3. STAR Panel Requests to the Stock Assessment Team (STAT-2)

Public Comments

(3:30 p.m.)

STAR Panel Discussion/Planning (as needed)

(4:00 p.m.)

Adjourn for the Day

(4:30 p.m.)

Tuesday, May 4, 2021 – 8:30 AM

Early Log-In to Resolve Connection Issues

(8:30 a.m.)

Responses to Panel Requests for Spiny Dogfish

(8:45 a.m.)

1. Presentation of Modeling Results — Vlada Gertseva

2. Further Discussion of Modeling

Results BREAK (10:00 – 10:15 a.m.)

3. Additional STAR Panel Requests to

STAT-1 LUNCH BREAK (11:30 a.m. – 1:00

p.m.)

Responses to Panel Requests for Dover Sole

(1:00 p.m.)

1. Presentation of Modeling Results — Chantel Wetzel & Aaron Berger
2. Further Discussion of Modeling

Results BREAK (2:15 – 2:30 p.m.)

3. Additional STAR Panel Requests to STAT-2

Public Comments

(3:30 p.m.)

STAR Panel Discussion/Planning (as needed)

(4:00 p.m.)

Adjourn for the Day

(4:30 p.m.)

Wednesday, May 5, 2021 – 8:30 AM

Early Log-In to Resolve Connection Issues

(8:30 a.m.)

Responses to Panel Requests for Spiny Dogfish

(8:45 a.m.)

1. Presentation of Modeling Results — Vlada Gertseva

2. Further Discussion of Modeling

Results BREAK (10:00 – 10:15 a.m.)

3. Additional STAR Panel Requests to

STAT-1 LUNCH BREAK (11:30 a.m. – 1:00

p.m.)

Responses to Panel Requests for Dover Sole

(1:00 p.m.)

1. Presentation of Modeling Results — Chantel Wetzel & Aaron Berger
2. Further Discussion of Modeling

Results BREAK (2:15 – 2:30 p.m.)

3. Additional STAR Panel Requests to STAT-2

Public Comments

(3:30 p.m.)

STAR Panel Discussion/Planning (as needed)

(4:00 p.m.)

Adjourn for the Day

(4:30 p.m.)

Thursday, May 6, 2021 – 8:30 AM

Early Log-In to Resolve Connection Issues

(8:30 a.m.)

Responses to Panel Requests for Spiny Dogfish

(8:45 a.m.)

1. Presentation of Modeling Results — Vlada Gertseva

2. Further Discussion of Modeling

Results BREAK (10:00 – 10:15 a.m.)

3. Agreement of a Preferred Model Between the STAR Panel and STAT-1
4. STAR Panel Requests for Model Runs for the Decision

Table LUNCH BREAK (11:30 a.m. – 1:00 p.m.)

Responses to Panel Requests for Dover Sole

(1:00 p.m.)

1. Presentation of Modeling Results — Chantel Wetzel & Aaron Berger
2. Further Discussion of Modeling

Results BREAK (2:15 – 2:30 p.m.)

3. Agreement of a Preferred Model Between the STAR Panel and STAT-2
4. STAR Panel Requests for Model Runs for the Decision Table

Public Comments

(3:30 p.m.)

STAR Panel Discussion/Planning (as needed)

(4:00 p.m.)

Adjourn for the Day

(4:30 p.m.)

Friday, May 7, 2021 – 8:30 AM

Early Log-In to Resolve Connection Issues

(8:30 a.m.)

Consideration of Remaining Issues

(8:45 a.m.)

1. Discussion of Proposed Base Models
2. Review Decision Tables for All

Assessments BREAK (10:00 – 10:15 a.m.)

3. Review Any Possible Disagreements from GMT, GAP, and PFMC Advisors
4. Identify Research and Data Needs

Public Comments

(11:00 a.m.)

LUNCH BREAK (11:30 a.m. – 1:00 p.m.)

Review Draft STAR Panel Report

(1:00 p.m.)

1. Discuss Deadlines for Report Submission
2. Review and Discuss Draft

Report BREAK (2:15 – 2:30

p.m.)

STAR Panel Discussion/Planning (as needed)

(2:30 p.m.)

STAR Panel Adjourns

(4:30 p.m.)

Annex 1: Format and Contents of CIE Independent Peer Review Report

1. The CIE independent report shall be prefaced with an Executive Summary providing a concise summary of the findings and recommendations, and specify whether the science reviewed is the best scientific information available.

2. The main body of the reviewer report shall consist of a Background, Description of the Individual Reviewer's Role in the Review Activities, Summary of Findings for each ToR in which the weaknesses and strengths are described, and Conclusions and Recommendations in accordance with the ToRs.

   a. Reviewers should describe in their own words the review activities completed during the panel review meeting, including providing a brief summary of findings, of the science, conclusions, and recommendations.

   b. Reviewers should discuss their independent views on each ToR even if these were consistent with those of other panelists, and especially where there were divergent views.

   c. Reviewers should elaborate on any points raised in the Summary Report that they feel might require further clarification.

   d. Reviewers shall provide a critique of the NMFS review process, including suggestions for improvements of both process and products.

   e. The CIE independent report shall be a stand-alone document for others to understand the weaknesses and strengths of the science reviewed, regardless of whether or not they read the summary report. The CIE independent report shall be an independent peer review of each ToRs, and shall not simply repeat the contents of the summary report.

3. The reviewer report shall include the following appendices:

   Appendix 1: Bibliography of materials provided for review
   Appendix 2: A copy of the CIE Statement of Work
   Appendix 3: Panel Membership or other pertinent information from the panel review meeting.

Appendix 3: Panel Membership or other pertinent information from the panel review meeting

The Panel was composed of two independently appointed Center for Independent Experts (CIE) reviewers (Dr. N. Cadigan, Canada; Dr. M. Cieri, US), an independent reviewer from Oregon State University (Dr. F. Caltabellotta) and an independent chair (Dr. Tien-Shui Tsou, Washington Department of Fish and Wildlife). Drs. Caltabellotta and Tsao are also members of the Pacific Fishery Management Council's (PFMC's) Science and Statistical Committee (SSC). The STAR Review Panel was supported and assisted by Mr. J. DeVore (PFMC), Ms. Whitney Roberts (Washington Department of Fish and Wildlife, PFMC Groundfish Management Team), and Mr. Gerry Richter (PFMC Groundfish Advisory Subpanel).

Assessment documents were prepared by stock assessment teams (STAT's) and presented by Dr. Chantel Wetzel of the Northwest Fisheries Science Center (NWFSC) and Dr. Aaron Berger (NWFSC) for Dover Sole, and by Dr. Vladlena Gertseva (NWFSC) and Dr. Ian Taylor (NWFSC) for Pacific Spiny Dogfish. They were assisted by Dr. John Wallace (NWFSC) and Dr. Sean E. Matson (National Marine Fisheries Service West Coast Region).

Appendix 4. Design-based methods versus model-based approaches

There is an extensive amount of statistical literature on design- versus model-based approaches in survey sampling research. There are also hybrid approaches, broadly referred to as model-assisted approaches (e.g., Särndal et al, 2003; Chen et al., 2004) that offer a good compromise between model-efficiency and the model-robustness of the design-based approach. I use notation similar to Chen et al. (2004) to describe these approaches.

Assume the survey area is divided into *N* distinct tow sites and that the catch variable of interest at site *i* is y_i and that the population average is $\bar{y}_N = N^{-1}\sum_{i=1}^{N} y_i$. Note that *N* is usually very large and it is impossible to sample every site. Assume that $n << N$ sites are sampled in a survey using a probability sampling design where the probability of sampling at site *i* is $\pi_i = \Pr(i \in s) > 0$ and the sample *s* are the *n* sites chosen to trawl at. A generic designed-based estimate of $\bar{y}_N$ is

$$\bar{y}_s = N^{-1}\sum_{i=1}^{n} y_i/\pi_i.$$

In statistics this is called the Horvitz-Thompson estimator. For example, if a simple random sampling design is used then $\pi_i = \frac{n}{N}$ and $\bar{y}_s = \bar{y}$ is the ordinary sample mean. If the population is divided into *H* strata and stratified simple random sampling is used with n_h samples in stratum $h = 1, \ldots, H$ then $\pi_{i\in h} = n_h/N_h$, $\bar{y}_s = N^{-1}\sum_{i=1}^{n} N_h\bar{y}_h$, and $\bar{y}_h$ is the ordinary sample mean for stratum *h*.

In fisheries surveys there will typically be measurement error in the catches, and the best we can hope for is that the trawl catch is unbiased for the local trawlable abundance at site *i* which I denote as μ_i; that is, $E_M(y_i) = \mu_i$. I assume the stochastic processes that generate the catch have some probability distribution function that is used for the model-based expectation E_M. I use the subscript D to denote design-based expectations where the average is with respect to all possible samples *s*. Typically the model will use auxiliary covariates (e.g., latitude, longitude, depth) that are known for all tow sites *1,…,N* and parameters that must be estimated using the sampled catches $\{y_i\}_{i\in s}$ to estimate the μ_i at all tow sites. Let $\hat{\mu}_i$ denote the estimate. The purely model-based estimate of $\mu_N = N^{-1}\sum_{i=1}^{N}\mu_i$ is

$$\hat{\mu}_N = N^{-1}\sum_{i=1}^{N}\hat{\mu}_i.$$

This is the VAST approach, where a model is used to predict μ_i at all tow sites. However, for various reasons the model may not always provide unbiased predictions, and this can create design-bias in $\hat{\mu}_N$ as an estimate of μ_N. Sometimes the bias can be severe (Chen et al., 2004).

An operational disadvantage of purely model-based approaches is that model assumptions must be appropriate and model estimation must be sufficiently reliable. This requires examining model diagnostics for each survey variable of interest, and typically there are many variables of interest for a species (i.e., number per tow, weight per tow, number per tow and length class, etc.) and many species of interest in fisheries surveys. It will usually be impractical to examine model goodness-of-fit for many variables. A single best model may not be apparent either, or model selection statistics such as AIC may guide us to the incorrect model (Thorson et al., 2021). Opsomer et al. (2007) referred to *generic* inference as the problem of making sensible estimates for many variables in a straightforward and internally consistent way, and they referred to

specific inference in which custom models are built for a few variables and the dataset at hand. I suggested that the human resource limitations typical of almost all fisheries science organizations that support stock assessment means that our focus should be on generic inference. There is therefore an understandable reluctance to specify statistical models for the behavior of all the variables of interest in the population (Breidt and Opsomer, 2017), whereas the design-based approach provides a simple and robust all-purpose statistical framework. A disadvantage of the design-based approach is that the resulting estimators can be inefficient, and sometimes dramatically so.

There are many model-assisted approaches that have been designed to provide design-unbiased estimates of μ_N (see Skinner and Wakefield, 2017; Breidt and Opsomer, 2017) but also improved efficiency compared to purely design-based estimators. Design consistency of model-assisted estimators is guaranteed, under very weak assumptions, and in particular consistency does not depend on strong modeling assumptions (Skinner and Wakefield, 2017). An intuitive approach is the difference estimator of μ_N,

$$\hat{\mu}_{Diff} = N^{-1}\sum_{i=1}^{N}\hat{\mu}_i + \sum_{i\in s}\frac{y_i - \hat{\mu}_i}{\pi_i} = \hat{\mu}_N - bias(\hat{\mu}_N),$$

where $-\sum_{i\in s}\frac{y_i-\hat{\mu}_i}{\pi_i}$ is a design-based estimate of the bias in $\hat{\mu}_N$ as an estimate of $\bar{y}_N$, and is an approximately design- and model-based estimate of the bias in $\hat{\mu}_N$ as an estimate of μ_N. This estimator will be approximately design-unbiased and is design-consistent regardless of any potential misspecification of the model. However, if the model-based $\hat{\mu}_i$ are highly correlated with the y_i then the design-variance of $\hat{\mu}_{Diff}$ will be much smaller than the design-based estimator, $\bar{y}_s$.

A problem for stock assessment is that $\hat{\mu}_{Diff}$ can be negative and for that and other reasons alternative model-assisted approaches have been proposed. For example, Liang et al. (2017) proposed Bayesian model calibration with the pseudo-empirical likelihood framework to produce improved Blue Crab (*Callinectes sapidus*) abundance indices. Model calibration approaches are an active area of research in survey sampling (e.g., Wu and Thompson, 2020).

As a first step, I think it will be useful to investigate if there is evidence of substantial VAST model bias. This only requires evaluating stratum size-weighted averages of the VAST ordinary raw residuals (observed minus model predicted). This is not an analysis of the VAST model goodness of fit, but rather just an evaluation if the VAST model predictions given unbiased predictions of trawl catches at sampled sites. This is the $bias(\hat{\mu}_N)$ term in the above equation. If the absolute average bias is large then additional and detailed examination of the VAST assumptions and estimation will be necessary. If VAST provides biased predictions of the trawl catches at the sample sites on average, then this casts doubt on the reliability of the VAST predictions for unsampled sites.

The standard designed-based estimator for a stratified random survey will be the same as a stratum-effects model-based approach, in which each stratum*year combination is a separate parameter in a statistical model. If the mle of the mean is the sample mean (i.e., Normal, Gamma, Poisson, Negative Binomial, delta-Gamma distributions) then the strata size-weighted average of model predictions will be the same as the design-based estimator. Hence, in the sense

of Firth and Bennett (1998), the stratum-effects model is design-consistent which is a desirable property.

Note that estimation of the measurement error variance is a problem with the stratum effects model when there are many strata and low sample sizes within strata. MLE's of variance parameters have a known bias that is not ignorable when the number of parameters is large relative to the sample size. This is a problem for statistical inferences about stock size (Cadigan, 2011). Also, the stratum-effects model cannot be used directly to interpolate trawlable densities in incomplete surveys in which not all strata are sampled in some years. Hence, I am not advocating for the stratum-effects model, but I just use this as an example of a desirable model property.

Breidt, F.J. and Opsomer, J.D., 2017. Model-assisted survey estimation with modern prediction techniques. Statistical Science, 32(2), pp.190-205.

Cadigan, N.G., 2011. Confidence intervals for trawlable abundance from stratified-random bottom trawl surveys. Canadian Journal of Fisheries and Aquatic Sciences, 68(5), pp.781-794.

Chen, J., Thompson, M.E. and Wu, C., 2004. Estimation of fish abundance indices based on scientific research trawl surveys. Biometrics, 60(1), pp.116-123.

Firth, D. and Bennett, K.E., 1998. Robust models in probability sampling. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60(1), pp.3-21.

Liang, D., Nesslage, G., Wilberg, M. and Miller, T., 2017. Bayesian calibration of blue crab (*Callinectes sapidus*) abundance indices based on probability surveys. Journal of Agricultural, Biological and Environmental Statistics, 22(4), pp.481-497.

Opsomer, J.D., Breidt, F.J., Moisen, G.G. and Kauermann, G., 2007. Model-assisted estimation of forest resources with generalized additive models. Journal of the American Statistical Association, 102(478), pp.400-409.

Särndal, C.E., Swensson, B. and Wretman, J., 2003. Model assisted survey sampling. Springer Science & Business Media.

Skinner, C. and Wakefield, J., 2017. Introduction to the design and analysis of complex survey data. Statistical Science, 32(2), pp.165-175.

Thorson, J.T., Cunningham, C.J., Jorgensen, E., Havron, A., Hulson, P.J.F., Monnahan, C.C. and von Szalay, P., 2021. The surprising sensitivity of index scale to delta-model assumptions: Recommendations for model-based index standardization. Fisheries Research, 233, p.105745.

Wu C., Thompson M.E., 2020. Calibration Weighting and Estimation. In: Sampling Theory and Practice. ICSA Book Series in Statistics. Springer, Cham. https://doi.org/10.1007/978-3-030-44246-0_6