Peer Review of Japanese Demersal Stocks in 2021

January 16, 2022

E.J. Dick, Ph.D.
National Marine Fisheries Service, Southwest Fisheries Science Center
Fisheries Ecology Division, Santa Cruz, California, USA
Email: edward.dick@noaa.gov

Background

The Fishery Research and Education Agency of Japan organized an independent peer review of demersal stock assessments in 2021. Invited peer reviewers for each stock included two Japanese experts and one external reviewer from NOAA Fisheries. Reviewers were asked to participate in the assessment review meetings and submit peer review reports (this document). Several documents were made available for consideration prior to the review panel meeting. These included stock assessment reports for the following three stocks:

1. Walleye Pollock (*Gadus chalcogrammus*), Japanese Pacific Stock
2. Walleye Pollock (*Gadus chalcogrammus*), Northern Sea of Japan Stock
3. Arabesque Greenling (*Pleurogrammus azonus*), Northern Hokkaido Stock

Other documents made available to the reviewers included responses to requests from the Committee of Stock Management Policy for both Walleye Pollock stocks, as well as a description of reference points for the Arabesque Greenling stock. The project manager of the peer review requested that the external reviewer prepare and submit pre-review questions to the assessment teams a month in advance of the meeting, to allow them time to organize their responses.

Reviewing system

The peer review was conducted November 17-18, 2021, via online meetings due to concerns related to COVID-19. While the virtual format provided an adequate review, a return to in-person meetings is recommended when possible. Interaction and communication between the reviewer and assessment team are limited when using a virtual meeting format, and the quality of the review process would be greatly improved by a return to in-person meetings.

Pre-review questions were requested and submitted (Appendix A) a month in advance of the meeting, and translation services were provided during the meeting itself. I appreciated having the opportunity to prepare questions in advance, as it would have been difficult for the stock assessment teams to prepare meaningful responses during the two-day review meeting. A large fraction of the presentations for each stock were devoted to answering the pre-review questions, which was an efficient use of time and provided an opportunity to explore additional questions as they arose.

On the first day of the meeting, introductions were followed by a presentation on the general framework for stock assessments in Japan. This included an overview of stocks currently managed using Total Allowable Catch (TAC), a flowchart of resource management relative to the assessment process, the roles of various research institutions, an overview of methods for setting MSY reference points, and a description of Harvest Control Rules (HCR). This framework is very similar to the management process used in the United States, and was therefore familiar. The material in this introductory presentation was quite dense, making reference to multiple agencies and stages of the review process. I recommend creating a short, written version of the presentation that can be distributed to external reviewers before the meeting, along with the terms of reference and stock assessment reports.

Two stocks of Walleye Pollock (Japanese Pacific and Northern Sea of Japan) were reviewed during the first meeting, followed by a single stock of Arabesque Greenling (Northern Hokkaido) on the second day. A review of three stocks in two days, including an introduction to the overall assessment framework, was an ambitious schedule. For comparison, stock assessment reviews in the western United States (i.e. via the Pacific Fishery Management Council) are conducted over five days for two stocks. If possible, additional time for review would provide an opportunity to better understand assessment model behavior, sensitivity to alternative assumptions, and identify recommendations for future research.

Stock assessment review report

1. Walleye Pollock, Japanese Pacific Stock (WP-JP)

   A. Determine whether the data used for stock assessment are adequate to understand the stock dynamics of the target species and represent the best scientific information available (BSIA).

      The modeling framework (VPA) used to assess the WP-JP stock requires information about annual catch at age. In response to pre-review questions (Appendix A), the assessment team did an excellent job of clarifying the process by which estimates of catch at age are derived. In particular, I found their descriptions of the catch reporting scheme, regional responsibilities by institute, and sampling methodology to be very helpful. The number of age composition samples by area, year, and fleet would be helpful in understanding the precision of age frequencies used to determine catch at age. I recommend that sampling information be made available to external reviewers prior to the review as part of the assessment report or as supplementary materials. Plots of age frequencies showed evidence of cohort progressions informing estimates of annual recruitment.

      Indices of abundance were developed using data from an offshore bottom trawl fleet (CPUE at age from Eastern Southwestern Hokkaido). Indices based on data from the anchored gillnet fleet (Southwestern Hokkaido), were derived from both skippers' notes and logbook data. Continued use of skippers' notes to develop CPUE indices is encouraged, as this data source contains valuable information for standardization of catch

and effort, as noted by the assessment team. The benefits of the logbook data include a longer time series. A fishery-independent index was also developed from acoustic and trawl survey data to inform estimates of recruitment during the last 3 years. It was suggested that the spatial coverage of this survey be expanded, into the Northern Territories if possible, because some dominant year classes (e.g. 2005, 2007) were not detected by the survey.

Based on the information provided prior to and during the review, these primary data sources meet the BSIA standard. However, potential areas for improvement were discussed during the review, and recommendations specific to these and other data types are included in later sections of this report.

B. Discuss whether the biological parameters used for stock assessment are appropriate.

The majority of discussion regarding biological parameters for the WP-JP stock focused on estimates of the natural mortality parameter (M). Estimates of M for this stock have been based on the method of Widrig (1954) since the 1995 assessment. Estimates based on alternative methods were prepared by the assessment team for comparison during the review, with results generally in the range of 0.2-0.3 for older age classes (3+), and a wider range of values for younger age classes (ages 0-2). It was noted that plans exist to directly estimate M for younger ages using the acoustic trawl method of Zwolinski and Demer (2013).

Since estimates of natural mortality affect estimates of stock size and fishing pressure, it is important to propagate uncertainty in M into management advice. This can be accomplished using a decision table framework, where alternative possible states of nature (e.g. models with different values of M) are used to evaluate a set of management actions (e.g. different catch levels) under each state of nature. M could also be estimated as a parameter in a statistical catch at age model, with or without a prior distribution developed from available data (e.g. maximum age, estimates for closely related stocks, etc.).

Maturity at age was also discussed during the review, with the majority of mature fish being ages 4+. The assessment team referred to the results of Hamatsu and Yabuki (2007), and although this study did not find conclusive evidence of density-dependent maturation in females, the authors identified possible sources of sampling bias that may have concealed the effect. To better understand the impact of maturity at age on assessment results, a sensitivity analysis is recommended in which alternative models are run using knife-edge age at maturity for ages 3 and 5, bracketing the age at 50% maturity in the base model. This analysis will help determine the sensitivity of model outcomes to differences in the assumed proportion of mature fish at age.

Growth (length at age and weight at age) was found to be density dependent in the study by Hamatsu and Yabuki (2007). The stock assessment report presents point estimates of length at age and weight at age used in the base model. It was not clear if interannual variability in size at age was accounted for in the model, or if the same average weights at

age were used in all years to convert estimates of numbers at age to biomass. I recommend that future assessment provide additional details regarding estimation of size at age (including data sources, sample design and sample sizes), whether or not growth is time-varying, and evidence (or lack thereof) for sexually dimorphic growth in the WP-JP stock.

C. Discuss whether the basic biological information such as distribution, migration pattern, and population are appropriate.

The WP-JP stock was defined along the Japanese Pacific coast, with several spawning areas and nursery grounds over the continental shelf. Movement patterns between the spawning grounds and nursery areas are inferred from both tagging studies and patterns in catch at age. There is some uncertainty regarding the proportion of the stock residing around the Northern Territories. Movement to/from this region and catches by the Russian fleet are not well understood due to data limitations, and the potential presence and influence of northern spawning grounds deserves further investigation.

The model assumes that expected recruitment is a function of the spawning stock biomass, i.e. that fecundity is proportional to the weight of mature females in the population. Many marine species, including some stocks of Walleye Pollock (Tanaka et al. 2016), have been shown to have a disproportionate increase in egg production with increasing female size. Kooka (2012) reported fecundity-length exponents greater than 3, suggesting that weight-specific fecundity (eggs per unit weight) increases with size. The assessment report should justify the assumption that egg production is proportional to female biomass, or otherwise model recruitment as a function of total egg production rather than mature female biomass.

D. Evaluate whether the stock assessment methodology is based on the most appropriate available study and performed analytically.

The stock assessment was conducted using a Ridge VPA framework (Okamura et al. 2017). Terminal Fs for ages 3-9 were determined by tuning the model to age-specific indices derived from the offshore trawl fleet, and aggregate abundance indices (skippers' notes and logbook) from the anchored gillnet fleet. Indices were weighted based on their fit to abundance or SSB (see comments, below, regarding the derivation of weights). Terminal Fs were estimated using maximum likelihood. A penalty term based on the square of F is included to minimize retrospective bias patterns.

The assessment team compared alternative weight parameters (lambda and eta) to the base case assumptions to illustrate the effect of the penalty term. Retrospective patterns were reduced in the base model relative to the lambda=0 option, as expected, and there was little change in the sigma values (describing fits to the indices) between the base case and the model with no penalty (lambda=0). This confirms that the penalty term in the base case model reduces retrospective bias without significantly degrading the fit to the tuning indices.

Recruitment in the last three years of the VPA model (2017-2019) were estimated using linear extrapolation based on a fitted relationship between the age-1 survey index and predicted age-1 abundance from the VPA. A comparison was made to methods used in previous assessments (i.e. assuming average recruitment based on recent years), but it was agreed that the estimates based on the survey were the best available data to inform recent recruitments.

E. Evaluate whether the data are treated statistically correctly.

Catches in the Northern Territories may bias assessment results if fish taken from those areas are part of the same biological stock. Continued collaboration with relevant agencies is encouraged to develop estimates of catch at age for both regions, if the current VPA framework is continued. Estimates of total catch with length-frequency and/or age-frequency samples could also be used to estimate stock dynamics using statistical catch at age models.

With respect to the catch at age estimation for the VPA model, I recommend further documentation of methods used to substitute information in unsampled strata, and development of a standardized report showing sample sizes and catch in weight by year, month, region, and fishery. Also, sampling of age compositions occurs at major ports within a region (e.g. Urawaka, Muroran), but it was not clear what fraction of total landings occurred at these ports relative to unsampled ports. If unsampled ports represent a non-trivial fraction of the landings, then occasional sampling of minor ports could be used to confirm that age compositions estimated from major ports are representative of the entire region.

Many details of the CPUE standardization process were not available in advance of the meeting. I recommend that documents with details of the standardization be provided to reviewers with the stock assessment report. It was noted that some indices were provided by another agency, but documentation of the methods is needed to allow for a detailed review, and could be provided as a separate document or included in the stock assessment report. Documentation of the GLM and delta-GLM approaches used for the skippers note index and offshore bottom trawl index would be useful, including sample sizes, model selection, and regression model diagnostics. Statistical methods for design-based indices should also be made available, and include estimates of uncertainty (e.g. confidence intervals) for each year.

The weighting of alternative abundance indices was described as a function of "the variance between the tuning indices and the relative abundance or SSB." If I understand correctly, this means that an index that is inconsistent with model results is down-weighted based on its poor fit to the SSB trend, which seems problematic. Weighting for indices should be a function of the uncertainty in the index itself (e.g. observation error). In this way, well-informed (precise) indices have greater influence on the model relative to imprecise indices. Estimates of uncertainty should be available from both the model-based (GLM and delta-GLM) and design-based indices, and could be used as weights in the likelihood functions.

Use of skippers' logs as a separate (independent) index, in addition to logbook data, should be evaluated to make sure that data are not being used twice, as this would violate the assumption of independence.

F. Evaluate whether the stock assessment result obtained from the input data and methodology used is appropriate.

The stock assessment results indicate a period of reduced recruitment from 2010-2015, with an increase in 2016. Estimated recruitments based on the survey (2017-2019) and used for forecasting purposes assume that 2017 recruitment was similar to 2016, with subsequent declines in recruitment from 2018-2019. Estimates of biomass and spawning biomass fluctuated between 800-1400 and 150-560 thousand mt, respectively.

Spawning stock biomass in 2019 was estimated at 302 thousand mt, based on an assumed natural morality rate of 0.25. Alternative values of M change the estimate from 200 thousand mt (M=0.05) to around 500 thousand mt (M=0.40). As noted by the stock assessment team, the value of M scales estimates of biomass, recruitment, and fishing pressure. The effect of M on reference points was also presented for alternative stock-recruitment curves and optimization methods (least squares vs. least absolute deviations). Results showed that MSY was generally lower under the hockey-stick model, except for extremely low values of M. Given the analysis of alternative M estimators, the base model assumption of M=0.25 for ages 3+ seems reasonable. However, as noted during the panel, exploration of methods to propagate uncertainty in M into forecasts and estimated probabilities of exceeding target and limit reference points is encouraged.

G. Evaluate the validity of methodology and result used for the future projection.

Assessment results (SSB and recruitment) were used to fit alternative stock-recruitment (SR) relationships. Each functional form for the SR relationship was also estimated using different optimization algorithms. Although the AICc model selection criterion identified a Ricker form as the best fit to the data, a hockey-stick (HS) SR relationship was chosen because declines in recruitment at large stock sizes were all thought to have occurred during a period of poor environmental conditions (2010-2015) that was not representative of average conditions. Selection of the HS model was completed outside of the review, but the choice seemed reasonable given the materials presented.

In a prior meeting, stakeholders requested a constant-catch management option, rather than the default harvest control rule (HCR). The assessment team presented a comparison of the default HCR to numerous constant-catch scenarios, ranging from 140-190 thousand metric tons. Expected SSB and catch associated with alternative "beta" values were presented for the forecast period (through 2031), along with associated probabilities of exceeding the target and limit reference points. Probabilities were calculated using simulated recruitments and assumption of independent, lognormal deviations from the base model stock-recruitment relationship (hockey-stick). Auto-correlation in recruitment was not considered.

During the review, there was some discussion of methods used to estimate parameters of the stock-recruitment relationships (e.g. least squares, least absolute deviation). Each of these assumes that SSB is known without error, and that all variability occurs in recruitment. Ludwig and Walters (1981) showed that it is beneficial to consider models that allow for uncertainty in both recruitment and SSB (i.e. an "errors in variables" approach) would change parameter estimates. This would require estimates of uncertainty in SSB, but could be explored by assuming a range of ratios between the variance in SSB and variance in recruitment. The current model selection approach identifies a single, "best" functional form for the stock-recruitment relationship, but this ignores uncertainty in the functional form. Another option to consider is a semi-parametric model that allows for uncertainty in the function itself (Munch et al. 2005). A third option is to estimate parameters of the SR relationship within the population dynamics model using a statistical catch at age framework.

2. Walleye Pollock, Northern Sea of Japan Stock (WP-NSJ)

A. Determine whether the data used for stock assessment are adequate to understand the stock dynamics of the target species and represent the best scientific information available (BSIA).

The modeling framework (VPA) used to assess the WP-NSJ stock requires information about annual catch at age. In response to pre-review questions, the assessment team noted that Russian catches are non-disclosure and may be composed of two Walleye Pollock stocks. Similarly, discard information was unavailable. If Russian catches and/or discard mortality of the WP-NSJ stock are a significant fraction of total fishing mortality, then assessment results will be biased because assessment models (both VPA and statistical catch at age) typically assume that catches are known precisely and without bias.

The spatial and temporal allocation of catch at age was presented, based on estimates created by the prefectural fisheries research institute in Hokkaido. Similar to the WP-JP stock, I recommend that sampling information, including sample sizes by area, year, and fleet, be made available to external reviewers prior to the review as part of the assessment report or as supplementary materials. As noted by the assessment team, this information is not currently included in stock assessment reports, but it could be made available as supplementary materials in a standardized format by the agency responsible for estimation of catch at age. Without details related to data collection in general, it is not possible (for external reviewers, in particular) to judge whether data are adequate to understand stock dynamics.

Indices of abundance for the WP-NSJ stock were developed using data from fishery-independent surveys, including an acoustic and bottom trawl survey of spawning biomass, age-0 (pelagic larvae and juveniles) spring survey, and age-1 acoustic and bottom trawl survey. The assessment team clarified that the indices were design-based,

rather than model-based, and that estimates of uncertainty were not available (see Section E, below, for additional comments on statistical properties of acoustic trawl surveys).

Based on the information provided during the peer review, these primary data sources appear to be the best available. Other recommendations, specific to individual data sources, are included in later sections of this report.

B. Discuss whether the biological parameters used for stock assessment are appropriate.

Several alternative methods for estimating the natural mortality parameter (M) were presented and compared to the base model, which used the same values as the WP-JP stock. Since it is difficult to estimate M, it is important to propagate uncertainty in M into management advice. As noted for the WP-JP stock, this can be accomplished using a decision table framework, where alternative possible states of nature (e.g. models with different values of M) are used to evaluate a set of management actions (e.g. different catch levels) under each state of nature. M could also be estimated as a parameter in a statistical catch at age model, with or without a prior distribution developed from available data (e.g. maximum age, estimates for closely related stocks, etc.).

Assumed values of length, weight, and maturity at age were briefly covered in the presentation. A sensitivity analysis is recommended in which alternative models are run using knife-edge age at maturity for ages 3 and 5, bracketing the age at 50% maturity in the base model. This analysis will help determine the sensitivity of model outcomes to differences in the assumed proportion of mature fish at age. As with the WP-JP stock, it was not clear if interannual variability in size at age was accounted for in the model, or if the same average weights at age were used in all years to convert estimates of numbers at age to biomass. I recommend that future assessment provide additional details regarding estimation of size at age (including data sources, sample design and sample sizes), whether or not growth is time-varying, and evidence (or lack thereof) for sexually dimorphic growth.

C. Discuss whether the basic biological information such as distribution, migration pattern, and population are appropriate.

The WP-NSJ stock was defined as having a central distribution off the west coast of Hokkaido, with three primary spawning areas and nursery grounds to the north in recent years. Movement patterns to/from the northern region and catches by the Russian fleet are not well understood due to data limitations, and the potential for stock connectivity between areas fished by Japanese and Russian fleets deserves further investigation.

The model assumes that expected recruitment is a function of the spawning stock biomass, i.e. that fecundity is proportional to the weight of mature females in the population. Many marine species, including some stocks of Walleye Pollock (Tanaka et al. 2016), have been shown to have a disproportionate increase in egg production with increasing female size. Kooka (2012) reported fecundity-length exponents greater than 3, suggesting that weight-specific fecundity (eggs per unit weight) increases with size. The

assessment report should justify the assumption that egg production is proportional to female biomass, or otherwise model recruitment as a function of total egg production rather than mature female biomass.

D. Evaluate whether the stock assessment methodology is based on the most appropriate available study and performed analytically.

The WP-NSJ stock assessment was conducted using a Ridge VPA framework (Okamura et al. 2017). Terminal Fs were estimated by tuning the model to an index of spawning biomass (mature fish ages 2-10+), an index of age-0 fish based on a survey of pelagic larvae and juveniles, and an index based on age-1 fish from a survey of juveniles (ages 0-2). Indices were weighted in the likelihood function, with greater weight assigned to the index of spawning biomass, as compared to the age-0 and age-1 indices (see comments, below, regarding the derivation of weights). Terminal Fs were estimated using maximum likelihood. A penalty term based on the square of F is included to minimize retrospective bias patterns.

The assessment team compared alternative values of the weight parameter (lambda) to the base case assumptions to illustrate the effect of the penalty term. Retrospective patterns were similar in the base model (lambda = 0.878) relative to the lambda=0 option, and fits to the indices were also similar based on visual inspection of residual plots. This confirms that the penalty term in the base case model reduces retrospective bias without significantly degrading the fit to the tuning indices. A pattern was noted whereby larger values of lambda (0.999) resulted in degraded fits to the indices and larger retrospective patterns, but there was not time during the review to fully discuss the issue.

Recruitment was modeled using a hockey stick (HS) functional form. Estimates of spawning biomass were all below the break point of the HS model, and suggest that the spawning stock has been below 25% of SB_{MSY} since roughly 2005. Fishing mortality rates were above target from 1990 through 2014, and have been near (slightly above or below) the target rate since 2015. Expected recruitment over the observed range of stock sizes was not sensitive to alternative stock recruitment models (Ricker, Beverton-Holt), or optimization algorithms (least squares vs. least absolute deviation). Variability in recruitment was quite high, with a log-scale standard deviation of 0.8.

E. Evaluate whether the data are treated statistically correctly.

Some comments regarding data treatment for the WP-NSJ stock are similar to the Japanese Pacific stock. With minor modifications, these topics are repeated in this section for the reader's convenience.

Russian catches in the west Sakhalin area may bias assessment results if fish taken from those areas are part of the WP-NSJ biological stock. Continued collaboration with relevant agencies is encouraged to develop estimates of catch at age for both regions, if the current VPA framework is continued. Estimates of total catch with length-frequency

and/or age-frequency samples could also be used to estimate stock dynamics using statistical catch at age models.

With respect to the catch at age estimation for the VPA model, I recommend further documentation of methods used to substitute information in unsampled strata, and development of a standardized report showing sample sizes and catch in weight by year, month, region, and fishery. Sampling of age compositions occurs at major ports within a region (e.g. Wakkanai, Otaru), but it was not clear what fraction of total landings occurred at these ports relative to unsampled, minor ports, if any. If unsampled ports represent a non-trivial fraction of the landings, then occasional sampling of minor ports could be used to confirm that age compositions estimated from major ports are representative of the entire region.

Many details related to the calculation of abundance indices were not available in advance of the meeting. I recommend that documents describing the amount of survey effort by area and year be provided to reviewers with the stock assessment report. Statistical methods for design-based indices should also be made available, as supporting documents if not part of the assessment report, and include estimates of uncertainty (e.g. confidence intervals) for each year. The assessment team noted that no estimates of uncertainty are available for the acoustic and trawl surveys. Methods to estimate precision of other acoustic trawl surveys have been developed (e.g. Stierhoff et al. 2020) and could potentially be adapted for use in the surveys of the WP-NJS stock.

The weighting of individual abundance indices was higher for the index of spawning biomass (given a weight of 10) compared to the age-0 and age-1 indices (each given a weight of 1). This implies trends in spawning biomass are thought to be more credible than trends in recruitment surveys, but is generally an ad-hoc approach to weighting. It was noted that this approach may be replaced with the methods used for the WP-JP stock. However, index weights could be derived as a function of the precision of each index (see comments related to index weights in the WP-JP stock section for more details).

F. Evaluate whether the stock assessment result obtained from the input data and methodology used is appropriate.

During the modeled time period, stock assessment estimates of biomass and spawning biomass declined to roughly 10% of their peak values around 2007-2008. Biomass declined subsequently due to poor recruitments from 2007-2009, but later increased to 154 thousand tons (56 of which is spawning biomass) by 2019.

As noted by the stock assessment team, the assumed value of M scales estimates of biomass, recruitment, and fishing pressure. The effect of using alternative M values on biomass, spawning biomass, and recruitment estimates for the 2019 fishing season was presented during the review. Recruitment showed the greatest sensitivity to changes in M. As noted during the panel, exploration of methods to propagate uncertainty in M into forecasts and estimated probabilities of exceeding target and limit reference points is encouraged.

G. Evaluate the validity of methodology and result used for the future projection.

Recruitments for future projections of the WP-NJS stock were assumed equal to the SR relationship with lognormal error, with the exception of some recent year classes. The years 2020 and 2021 were assumed to have above-average recruitment, based on available indices of recruitment. A comparison of results based on the expected recruitment from the SR relationship is recommended, as few other data sources are currently available to verify the strength of these recent year classes. If forecasts are highly sensitive to the assumed recruitment values, uncertainty in the strength of these year classes could be identified as a major axis of uncertainty in a decision table framework.

A linear relationship was estimated for predicted recruitments from the WP-JP stock assessment and the recruitment index for that stock. It would be useful to see a similar plot and relationship for the WP-NSJ stock, to better understand the relationship between survey observations and VPA estimates. Expected values from the linear model could also be used to inform recruitment in recent years, similar to the methods used for the WP-JP stock.

Assessment results (SSB and recruitment) were used to fit alternative stock-recruitment (SR) relationships. Each functional form for the SR relationship was also estimated using different optimization algorithms. AICc model selection did not distinguish between alternative forms, and a hockey-stick (HS) SR relationship was chosen. Selection of the HS model was completed outside of the review, but the choice seemed reasonable given the materials presented.

In a prior meeting, stakeholders requested a constant-catch management option, rather than the default harvest control rule (HCR). The assessment team presented a comparison of the default HCR to numerous constant-catch scenarios, ranging from 140-190 thousand metric tons. Expected SSB and catch associated with alternative “beta” values were presented for the forecast period (through 2031), along with associated probabilities of exceeding the target and limit reference points. Probabilities were calculated using simulated recruitments and assumption of independent, lognormal deviations from the base model stock-recruitment relationship (hockey-stick). Auto-correlation in recruitment was not considered.

As with the WP-JP stock, I recommend investigation of methods to estimate parameters of the stock-recruitment relationships that do not assume that SSB is known without error. Ludwig and Walters (1981) showed that it is beneficial to consider models that allow for uncertainty in both recruitment and SSB (i.e. an “errors in variables” approach). This would require estimates of uncertainty in SSB, but could be explored by assuming a range of ratios between the variance in SSB and variance in recruitment. Another option is to estimate parameters of the SR relationship within the population dynamics model using a statistical catch at age framework.

3. Arabesque Greenling, Northern Hokkaido Stock (AG)

A. Determine whether the data used for stock assessment are adequate to understand the stock dynamics of the target species and represent the best scientific information available (BSIA).

The modeling framework (VPA) used to assess the AG stock requires information about annual catch at age. I found the description of the sampling methodology in the presentation (slides 48-50) to be very helpful, and recommend that future assessments provide similar information to reviewers unfamiliar with the catch sampling process. The presentation made reference to two publications by Takashima, and perhaps distribution of these references in advance of the review would be sufficient. Specifically, the number of age composition samples by area, year, and fleet would be helpful in understanding the precision of age frequencies used to determine catch at age. Plots of age frequencies showing evidence of cohort progressions would also be useful.

The AG assessment was tuned to a single index of abundance (CPUE for the offshore bottom Danish seine fishery, 2005-2019). The area-weighted index was standardized using a GLM with year, month, area, and two-way interaction terms with year. The use of area weights is consistent with treatment of CPUE as a relative measure of density, which is converted to an index of abundance via the area weights. This approach is preferable to an unweighted index, which assumes CPUE is a direct measure of abundance. Detailed comments and recommendations regarding the standardization process are in Section E, below.

Based on the information provided prior to and during the review, these primary data sources meet the BSIA standard. However, potential areas for improvement were discussed during the review, and recommendations specific to these and other data types are included in later sections of this report.

B. Discuss whether the biological parameters used for stock assessment are appropriate.

As requested in the pre-review questions (Appendix A), the AG stock assessment team provided results of a sensitivity analysis in which the base model estimate of M was halved and doubled. As expected this had the effect of scaling the estimated biomass trend, with biomass in 2019 decreasing by 13% when M was half the base case value, and increasing by 36% when M was double the base case. Estimates of 2019 spawning biomass were less variable, decreasing by 9% and increasing by 24% over the same range of M values. Unlike the Walleye Pollock assessments, natural mortality was assumed constant with respect to age, fixed at a value of M=0.295.

Since estimates of natural mortality affect estimates of stock size and fishing pressure, it is important to propagate uncertainty in M into management advice. This can be accomplished using a decision table framework, where alternative possible states of nature (e.g. models with different values of M) are used to evaluate a set of management

actions (e.g. different catch levels) under each state of nature. M could also be estimated as a parameter in a statistical catch at age model, with or without a prior distribution developed from available data (e.g. maximum age, estimates for closely related stocks, etc.). The AG assessment took a step towards a decision table, providing forecasts based on different assumptions about recruitment strength. If both recruitment and M were considered to be important sources of uncertainty, then models with low M/low recruitment and high M/high recruitment could be compared to the base model to provide a range of possible states of nature.

Maturity at age was also discussed during the review, and the majority of age-1 fish are assumed to be mature, and all age-2+ fish are assumed mature. A recent study found evidence of density-dependent maturation, with declines in the fraction of mature age-1 fish observed for larger year classes. Most age-structured assessment models (VPA and statistical catch at age) only account for density dependence via the stock-recruitment relationship, but examination of population dynamics with density-dependent maturity would be an appropriate topic of future research for this stock.

The stock assessment report presents point estimates of length at age and weight at age used in the base model. Although sex-specific equations were provided for length at age, the model appears to be a single-sex model. This assumes that both males and females are equally vulnerable to fishing gear regardless of differences in size at age. For this stock, the differences in size at age may be small enough to ignore, but I recommend that future assessment provide additional details regarding estimation of size at age (including data sources, sample design and sample sizes), whether or not growth is time-varying, and justification of the use of a single-sex model when used for assessment purposes.

C. Discuss whether the basic biological information such as distribution, migration pattern, and population are appropriate.

Fishing grounds for the AG stock were identified along the western and northern coast of Hokkaido, with spawning areas around Rishiri and Rebun Islands and Musashi bank. Adults are thought to remain near the spawning grounds. Insufficient data are available to inform patterns of egg or larval dispersal between areas around Hokkaido and more northerly waters. The southern Hokkaido stock has spawning grounds near the assessed stock, and connectivity patterns should be investigated and/or joint assessment considered, especially if trends in biomass and recruitment are found to be similar.

The model assumes that expected recruitment is a function of the spawning stock biomass, i.e. that fecundity is proportional to the weight of mature females in the population. Many marine species have been shown to have a disproportionate increase in egg production with increasing female size. Spawning frequency (number of broods per season) may also increase with size/age of females. The assessment report should justify the assumption that egg production is proportional to female biomass, or otherwise model recruitment as a function of total egg production rather than mature female biomass.

D. Evaluate whether the stock assessment methodology is based on the most appropriate available study and performed analytically.

The AG stock assessment was conducted using a Ridge VPA framework (Okamura et al. 2017). Due to recent changes in the selectivity of the coastal and offshore fisheries (avoidance of age-0 juveniles), terminal Fs were determined by tuning the model to such that retrospective patterns were minimized in the older age classes. This was accomplished through the use of a weighted sum (weight = lambda) comprised of a penalty term based on the square of F, plus a goodness of fit term based on the sum of squared errors between the bottom trawl index and the model's predictions. Relative to the Walleye Pollock assessments, this model had larger retrospective patterns.

The assessment team compared alternative weight parameter (lambda) values to the base case assumption (lambda = 0.09) to show the influence of the penalty term on model results. There was little change in the fit to the index between the base case and the model with no penalty (lambda=0). This confirms that the penalty term in the base case model reduces retrospective bias without significantly degrading the fit to the tuning indices.

E. Evaluate whether the data are treated statistically correctly.

With respect to the catch at age estimation for the VPA model, I recommend further documentation of methods used to substitute information in unsampled strata, and development of a standardized report showing sample sizes and catch in weight by year, month, region, and fishery. If sampling of age compositions only occurs at major ports within a region identify the fraction of total catch landed at major ports relative to unsampled ports. If unsampled ports represent a non-trivial fraction of the landings, then occasional sampling of minor ports could be used to confirm that age compositions estimated from major ports are representative of the entire region.

Some details of the CPUE standardization process were not available in advance of the meeting. I recommend that documents with additional details of the standardization be provided to reviewers prior to the peer review meeting. After revisiting the assessment report, it was not clear to me why the expected value of the year/area interaction terms was used for the standardization. In theory, the predicted CPUE in each year/area combination could be multiplied by the area, then summed across areas within each year to derive the index. Further documentation of the GLM would be useful, including sample sizes by year and area, model selection, and regression model diagnostics. The assessment team provided some diagnostics (a QQ-plot and analysis of deviance table) and model selection details during the review, which was greatly appreciated.

Weights for abundance indices were not used in the AG assessment because only a single index was used to for tuning. However, even within a single index, different years may vary in precision (e.g. years with small sample sizes due to poor weather may be imprecise). Standard errors associated with each year can be estimated from the GLM standardization process, and used as year-specific variances (weights) within the index likelihood.

F. Evaluate whether the stock assessment result obtained from the input data and methodology used is appropriate.

A stock recruitment relationship for the AG stock was proposed temporarily during a meeting of the Research Institute in April 2019. SSB and recruitment estimates available at that time were used to fit alternative stock-recruitment (SR) relationships. Each functional form for the SR relationship was also estimated using different optimization algorithms. Model selection via AICc identified a hockey stick functional form as the best fit to the data, estimated by minimization of absolute deviations. Although selection of the HS model was completed prior to the review, but the choice seemed reasonable given the materials presented.

The stock assessment results indicate a period of reduced recruitment beginning in 2010, resulting in lower biomass estimates since 2011. Estimated recruitments for 2017 and 2019 are thought to be above average, but forecasts accounted for the possibility of continued negative recruitment residuals in addition to a scenario based on expected recruitment with lognormal errors (see additional comments regarding future projections, for the AG stock in Section G, below). Spawning stock biomass in 2019 was estimated at 24 thousand mt, based on an assumed natural morality rate of 0.295. The base model results indicate that the stock has been below the target reference point since 2002, but above the proposed fishing ban level.

As noted by the stock assessment team, the value of M scales estimates of biomass, recruitment, and fishing pressure. Alternative values of M change the estimate from 20 thousand mt (M=0.1495) to around 30 thousand mt (M=0.59). Exploration of methods to propagate uncertainty in M into forecasts and estimated probabilities of exceeding target and limit reference points is encouraged.

G. Evaluate the validity of methodology and result used for the future projection.

Future projections were based on forward projection of the cohort analysis. Future recruitments (through 2031) were generated from the hockey stick model with lognormal deviations. Average body weight at age has been shown to decline with increasing abundance for this stock (Appendix 1 of the stock assessment report). Future projections accounted for this by generating estimates from a fitted relationship (with observation error).

Recent recruitment for this stock has been below the expected values (negative residuals). To account for the possibility of continued low recruitments (an alternative state of nature), the assessment team provided future projections based on resampled recruitment residuals that included the recent observations of poor recruitment.

Another point made by the stock assessment team, related to future projections, was that "the prediction value of catch and spawning biomass for 2021 will heavily depend on the assumption of the fishing mortality of age 1 fish in 2020." The assessment team

developed a methodology whereby uncertainty in F for 2020 can be propagated into future projections. Proper characterization of uncertainty in future projections, such as the approach taken here, is critical given the emphasis on probabilities of specific outcomes in the management system.

The methods used to estimate parameters of the stock-recruitment relationships (e.g. least squares, least absolute deviation) assume that SSB is known without error, and that all variability occurs in recruitment. Ludwig and Walters (1981) showed that it is beneficial to consider models that allow for uncertainty in both recruitment and SSB (i.e. an "errors in variables" approach) would change parameter estimates. This would require estimates of uncertainty in SSB, but could be explored by assuming a range of ratios between the variance in SSB and variance in recruitment. Using a statistical catch at age approach, it may be possible to estimate parameters of the SR relationship within the population dynamics model.

H. Other Comments

The format used to illustrate how alternative lambda values affect model fit to the indices was very helpful (e.g. slides 14 & 15), and I recommend that future assessments use a similar format. Ideally, the plots of predicted vs. observed trends in CPUE (or survey biomass) could be added to the residual plots and sigma values provided for the Walleye Pollock assessments. Once annual estimates of standard error are available from the GLM, those could be included in the plots of predicted versus observed trends, as well.

I appreciated the effort by the Arabesque Greenling team to included alternative states of nature (levels of recruitment) in their presentation of future projection results. This is similar to the "decision table" format mentioned previously, in which a set of catch scenarios (management options) is applied to multiple models, each with different assumptions about the true state of nature (e.g. different M values, uncertainty in F for 2020, and/or levels of recent recruitment).

Overall Comments

Overall, the quality of the three assessments for demersal stocks examined during the review, as well as the presentations given by the assessment teams, were excellent. I appreciate and recognize that the assessment teams spent considerable time and effort addressing numerous pre-review questions, and found their answers to be informative and reasonable. The ridge VPA modeling framework, although less familiar to me than the statistical catch at age models used in the United States, is an appropriate analytical framework for the available data.

Several agencies were involved in the preparation of data used to inform the stock assessments (e.g., estimates of catch at age, fishery-dependent CPUE indices, fishery-independent surveys, biological parameters). Since the quality of assessment model results depends on the quality of the data inputs, I recommend that standardized data reports be developed (if not already available) for each data type, including details of the sampling methodology and/or survey

design, sample sizes, design-based and model-based estimators, and estimates of uncertainty where possible. This information should be distributed in advance of the peer review meeting, either as part of the assessment document, or as supplemental materials. Additional details related to data collection and processing would improve reviewers' ability to identify potential issues and provide recommendations for improvements to the assessment process.

Since management advice associated with each assessment includes estimated probabilities of specific outcomes (e.g. SSB exceeding the target reference point), characterization of uncertainty is a key component of the management framework. The reviewed assessments currently account for recruitment variability in future projections, but do not propagate parameter uncertainty (e.g. parameters of the stock-recruitment relationship, growth, natural mortality) into forecasts. Alternative modeling frameworks (e.g. statistical catch at age) have the potential to incorporate multiple sources of uncertainty in model results, which can then be propagated into future projections.

Future projections for each assessment were conducted for a range of management actions (alternative catch scenarios), but usually based on a single 'best' model (the base case). In order to understand the implications of uncertainty in model outcomes, I recommend applying the same set of management actions to models representing alternative possible states of nature (i.e. a "decision table" framework). This requires identification of a "major axis" of uncertainty, i.e. a factor or set of factors that have the greatest impact on assessment results. For example, alternative values of natural mortality and/or strength of recent recruitments could bracket a range of outcomes around the base case results. This approach provides managers with information about potential outcomes associated with alternative management actions that better characterize uncertainty in model outcomes.

The method for defining reference points in Japan is different from other management systems that I've encountered. It begins with an estimate of MSY and the biomass that produces MSY (target SB = SB_{MSY}). The spawning biomass estimates for the limit and fishing ban reference points are then identified based on percentages of MSY (e.g. the biomass that produces 60% of MSY is the limit, and the biomass producing 10% of MSY is the fishing ban). Other management systems define the limit and ban reference points as percentages of SB_{MSY} (e.g., the SB_{limit} = X% of SB_{MSY} and the SB_{ban} = Y% of SB_{MSY}). Yields are then calculated associated with these biomass levels. It would be useful to evaluate the performance of these two alternatives using a Management Strategy Evaluation (closed loop simulation) framework, to better understand the trade-offs.

Finally, I would like to thank Dr. Nakano, Dr. Iwasaki, Dr. Manabe, members of the Secretariat, and the stock assessment teams for the opportunity to participate in the peer review process. It has been very educational and enjoyable for me, and I hope that my participation and comments have been useful.

References

Hamatsu, T. and K. Yabuki. 2007. Density effects on length at maturity of walleye pollock *Theragra chalcogramma* off the Pacific coast of northern Japan in the 1990s. Fisheries Science 73: 87–97.

Kooka, K. 2012. Life-history traits of walleye pollock, *Theragra chalcogramma*, in the northeastern Japan Sea during early to mid 1990s. Fisheries Research 113: 35-44.

Ludwig, D. and C. Walters. 1981. Measurement errors and uncertainty in parameter estimates for stock and recruitment. Canadian Journal of Fisheries and Aquatic Sciences 38: 711-720.

Munch, S., A. Kottas, and M. Mangel. 2005. Bayesian nonparametric analysis of stock–recruitment relationships. Canadian Journal of Fisheries and Aquatic Sciences 62: 1808–1821.

Okamura, H., Y. Yamashita, and M. Ichinokawa. 2017. Ridge virtual population analysis to reduce the instability of fishing mortalities in the terminal year. ICES Journal of Marine Science 74(9): 2427–2436.

Stierhoff, K., J. Zwolinski, and D. Demer. 2020. Distribution, biomass, and demography of coastal pelagic fishes in the California Current Ecosystem during summer 2019 based on acoustic-trawl sampling. U.S. Department of Commerce, NOAA Technical Memorandum NMFS-SWFSC-626. https://doi.org/10.25923/nghv-7c40

Tanaka, H., T. Hamatsu, and K. Mori. 2017. Comparison of potential fecundity models for walleye pollock *Gadus chalcogrammus* in the Pacific waters off Hokkaido, Japan. Journal of Fish Biology 90(1): 236-248.

Widrig, T. 1954. Method of estimating fish populations, with application to Pacific sardine. Fish. Bull. U.S., 56, 141-166.

Zwolinski, J. and D. Demer. 2013. Measurements of natural mortality for Pacific sardine (*Sardinops sagax*). ICES Journal of Marine Science 70(7): 1408–1415.

Appendix A

Pre-review questions regarding the stock assessments of demersal species in 2020

E.J. Dick, Ph.D.
NOAA Fisheries
Southwest Fisheries Science Center
October 17, 2021

Several documents were made available for consideration prior to the panel meeting. These included stock assessments of the following three species as of 2020:

1. Walleye Pollock, Northern Sea of Japan Stock (WP-NSJ)
2. Walleye Pollock, Japanese Pacific Stock (WP-JP)
3. Arabesque Greenling (atka mackerel), Northern Hokkaido Stock (AG)

Other documents included responses to requests from the Committee of Stock Management Policy for both Pollock stocks, as well as a description of reference points for the Greenling stock. The purpose of this document is to provide questions based on the available documentation, such that the assessment teams can address these questions during the panel meeting scheduled for November 17-18, 2021.

Having read the documents provided for the panel meeting, I believe that a two-day schedule is an ambitious timeline for a review of three stock assessments. There are many details related to data inputs, modeling choices, reference points, and forward projections, and any one of these could consume a day of review. However, my experience is mainly with demersal stocks managed by NOAA Fisheries in the western United States, and in our system a review of two stocks lasts five days. Given the relatively condensed timeline, I will do my best to prioritize comments and identify topics that I see as major issues.

<u>Questions relevant to all demersal stock assessments</u>

1. Clarify whether catches used in each assessment represent all sources of catch taken from the assessed area. Are catches by all foreign fisheries accounted for, or are there other sources of mortality (e.g., discarded catch), that could bias the total catch estimates? The VPA model assumes that catches are known exactly, so sources of bias in the catch estimates should be minimized to avoid bias in derived reference points and projections.
2. Provide additional details regarding the sampling programs that were used to estimate catch at age. In addition to the total catch (question #1, above), the method used to allocate catches to individual age classes in a given year should be documented in the assessment or in an appendix to the assessment. Sample sizes and average age by major fleet, year, and area would be useful.

3. Describe how indices of abundance are calculated, and provide tables of sample size (number of hauls, trips, etc.) by year and area for each index. Are survey indices derived from design-based estimators, or are model-based methods (e.g., generalized linear models) used to derive annual estimates? If model-based methods are used, describe the model selection procedure and diagnostics for the final model.
4. The Ridge VPA defines a penalty term (the square of F) with a weight (λ) that is chosen to minimize retrospective bias. The assessments provide plots showing the retrospective bias given the chosen weight (e.g., Appendix Figure 2-2 on page 32 of the WP-NSJ assessment). However, it would be helpful to see how fits to each abundance index change for alternative values of λ. I suggest plotting fits to each index using three values of λ: 1) the λ that minimizes the retrospective bias (this is already done), 2) λ=0, and 3) λ=0.99. The purpose of this is to show how fits to the abundance indices are affected by minimization of the retrospective bias.
5. Plots of spawning biomass versus time and associated reference points for reasonable alternative values of M (e.g., upper and lower bounds) would be useful to help communicate the implications of error in this parameter on management advice. Methods for deriving M are inconsistent among stocks. Standardization of estimation methods is preferred if no specific reasons are given for using a particular method.

<u>Questions specific to the assessment of Walleye Pollock, Northern Sea of Japan Stock in 2020</u>

6. Provide justification for the weights (W_k) given to each index of abundance. For example, the WP-NSJ assessment assigns a weight of 10 to the spawning biomass index, and a weight of 1 to the recruitment index.
7. Are uncertainty estimates available for the annual indices of spawning biomass and recruitment? The years and point estimates are provided in Appendix Tables 4-1 and 4-2, but standard errors of the annual survey estimates would help interpret perceived changes over time, and could potentially be used as relative weights in the tuning process.
8. The natural mortality rate (M) is assumed to be 0.25 per year for ages 3 and above, and 0.3 for age 2 fish. Please provide a rationale for these values, including any ageing studies done for the assessment or from the literature, comparisons with similar stocks, and methods for estimation of M. The authors indicate that individuals of age 10 and above have been caught. Based on recent meta-analyses (e.g. Then 2014, Hamel 2015) M estimates based on $\sim 5/A_{max}$ are in common use, which would suggest a value closer to M=0.5. The value in the assessment is more consistent with a maximum age of 25 years. Regardless, since M is not known with precision and affects estimates of F (and therefore abundance), model runs with alternative values for M would help managers understand risks associated with uncertainty in M.
9. I would appreciate an explanation of Figure 4-2, the changes in age-0 and age-1 fish from the recruitment survey. I expected a large cohort of age-0 fish (such as 2006) to appear as a large cohort of age-1 fish in 2007, but maybe I am misinterpreting the figures. My understanding is that the abundance of age-1 fish is used as the index for age-2 fish in the tuning procedure. If cohort abundance does not carry through to successive years, then the age-1 index may not be representative of age-2 fish that 'recruit' to the model.

Questions specific to the assessment of Walleye Pollock, Japanese Pacific Stock in 2020

10. Model selection based on AICc (Appendix Table 10-1) seemed to favor a Ricker stock-recruitment relationship using both optimization methods. Alternative stock-recruitment models may produce very different management reference points. Since the harvest control rule includes a linear reduction in F below the SB_{limit}, it was not clear why the hockey-stick recruitment model was needed to reduce risks associated with potential future declines in abundance.
11. In the WP-NJS assessment, unequal weights (W_k) given to each index of abundance. This assessment did not describe any weighting of different indices in the negative log-likelihood. Were all of the abundance indices equally weighted?
12. Does the weighting factor (α) in equation 8 retain a value of 20 in the final model? It seems that the chosen value will affect the influence of the penalty term. Fits to the abundance indices should be plotted for alternative weightings on the penalty term (see question #4). Minimizing the retrospective pattern by emphasizing the penalty term may cause a significant degradation of fit to the abundance indices.
13. Are uncertainty estimates available for the annual abundance indices? Standard errors of the annual survey estimates could potentially be used as relative weights in the tuning process, and otherwise help with interpretation of trends over time and/or interannual variability.
14. The natural mortality rate (M) is assumed to be 0.25 per year for ages 3 and above, with larger values for younger ages. Please describe any ageing studies done for the assessment or from the literature, and comparisons with similar stocks. I am not very familiar with Widrig (1954), but the authors indicate that individuals of age 20 are caught, although infrequently. Based on recent meta-analyses (e.g. Then 2014, Hamel 2015) M estimates based on $\sim 5/A_{max}$ are in common use, which is roughly consistent with the assumed value for age 3+ fish. However, since M is not known with precision and affects estimates of F (and therefore abundance), model runs with alternative values for M would help managers understand risks associated with uncertainty in M.
15. In this assessment, recent recruitments are informed by an age-1 survey, whereas previous assessments used an average of recent recruitments. A comparison of the two approaches would be welcome, given the recent change. Another alternative would be to assume recruitments were equal to the expected value from the assumed stock-recruitment relationship. The authors note that the survey did not detect the 2005 and 2007 year-classes, but that they were considered ‘dominant’ cohorts. A brief discussion of this inconsistency would be appreciated.

Questions specific to the assessment of Arabesque Greenling, Northern Hokkaido Stock in 2020

16. There was no discussion of alternative stock-recruitment relationships or model selection procedure for this stock. Alternative stock-recruitment models may produce very different management reference points, and an exploration of alternative models seems warranted to better characterize uncertainty in stock projections and reference points.

17. In the WP-NJS assessment, unequal weights (W_k) given to each index of abundance. This assessment did not describe any weighting of different indices in the negative log-likelihood. Were all of the abundance indices equally weighted?
18. Similar to the other assessments, a penalty term (the square of F) was weighted by λ, but I was unable to find the final λ value in the report. Fits to the abundance indices should be plotted for alternative weightings on the penalty term (see question #4). Minimizing the retrospective pattern by emphasizing the penalty term may cause a degradation of fit to the abundance indices as λ approaches a value of 1.
19. Are uncertainty estimates available for the annual abundance indices? Standard errors of the annual survey estimates could potentially be used as relative weights for individual years in the tuning process, and otherwise help with interpretation of trends over time and/or interannual variability. The offshore bottom trawl CPUE was standardized using a generalized linear model (GLM), so estimates of uncertainty for the year effects can be calculated from the GLM outputs. These could be used when calculating the negative log-likelihood in the tuning procedure, but it was not clear whether the authors took this approach or not. Diagnostics for the linear model (e.g. selection of covariates, residual plots, etc.) would be helpful to better understand the quality of the fit.
20. The natural mortality rate (M) is assumed to be 0.295 per year for all ages. Based on the reported longevity (8 or 9 years), M estimates based on $\sim 5/A_{max}$ would be closer to 0.5 per year. However, since M is not known with precision and affects estimates of F (and therefore abundance), model runs with alternative values for M would help managers understand risks associated with uncertainty in M.