ISC/22/ALBWG-01/01

Re-examining the Model Diagnostics of the 2020 North Pacific Albacore Tuna Stock Assessment

Steven L. H. Teo[1] and Carolina V. Minte-Vera[2]

1 NOAA Fisheries, Southwest Fisheries Science Center, La Jolla, California, USA

2 Inter-American Tropical Tuna Commission, La Jolla, California, USA

Email: steve.teo@noaa.gov

This working paper was submitted to the ISC Albacore Working Group Stock Assessment Improvements Workshop, 9-12 May 2022 held by webinar

Abstract
The ALBWG used a suite of model diagnostics to assess potential issues associated with convergence, model structure, parameter mis-specification, and data conflicts in the 2020 base case model. Some recent work on assessment model diagnostics and a workshop on model diagnostics by the Center for the Advancement of Population Assessment Methodology (CAPAM) in 2022, inspired us to re-examine the model diagnostics of the 2020 NPALB assessment. In this paper, we re-examine the R0 likelihood profiles and estimate the prediction skill of the 2020 base case model through hindcasting. First, a new set of R0 likelihood profiles were performed with recruitment deviates that were forced to sum to zero. Second, we examine the prediction skill of the 2020 base case model by hindcasting the model and calculating the mean absolute scaled error (MASE) for the primary abundance index of the model. The likelihood profiles of $\log(R_0)$ from the 2020 assessment and the current study, where the estimated recruitment deviations summed to zero, provided largely consistent information about the overall population scale of the base case model but with fleet-specific differences in the relative importance of the size composition data. Our interpretation of these results are that R0 profiles remain useful for identifying consistency (or lack thereof) between data sources on the estimated population scale. However, the use of R0 profiles to identify the importance of specific fleets would need further work because the recruitment deviations can interact with different data sources in multiple ways, which can result in different scales of misfits. The MASE scores indicated that the 2020 base case model had approximately 16% and 10% improvements in forecast accuracy of the F09 abundance index over horizons of 1 and 2 years, respectively, as compared to naive forecasts. However, the MASE for 3 or more years were >1, which indicated that forecasts of 3 or more years did not exhibit any prediction skill over naive forecasts. Given these results, 2020 base case model has sufficient skill to estimate and forecast current stock status with a horizon of 2 years. However, it may be useful to update the F09 abundance index during the intervening years to monitor the relative status of the spawning stock biomass.

Introduction
Stock assessments of the North Pacific stock of albacore tuna (NPALB) are regularly conducted by the Albacore Working Group (ALBWG) of the International Scientific Committee for Tuna and Tuna-like species in the North Pacific Ocean (ISC). The most recent NPALB stock assessment was conducted by the ALBWG in 2020 (ALBWG 2020) using the Stock Synthesis (SS) modeling platform (Methot and Wetzel 2013).

The ALBWG used a suite of model diagnostics to assess potential issues associated with convergence, model structure, parameter mis-specification, and data conflicts in the 2020 base case model. These model diagnostics included: 1) model convergence tests, 2) Age-Structured Production Model (ASPM) diagnostic (Maunder and Piner 2015), 3) likelihood profiles over the estimated unfished level of recruitment (R0) (Lee et al. 2014), 4) residual analysis, and 5) retrospective analysis (Mohn 1999). Based on these diagnostics, the ALBWG concluded that the base case model was able to estimate the stock production function and the effect of fishing on the abundance of the NPALB stock. Similar to the 2017 assessment, the link between catch-at-age and the abundance index adds confidence to the data used and the results of the assessment. Due to the moderate exploitation levels relative to stock productivity, the production function was weakly informative about the NPALB stock size, resulting in asymmetric uncertainty in the stock's absolute scale, with more uncertainty in the upper limit of the stock than the lower limit.

Some recent work on assessment model diagnostics (Carvalho et al. 2021; Kell et al. 2021) and a workshop on model diagnostics (http://www.capamresearch.org/Diagnostics-Workshop) by the Center for the Advancement of Population Assessment Methodology (CAPAM) in 2022, inspired us to re-

examine the model diagnostics of the 2020 NPALB assessment. Importantly, the attendees at the model diagnostics workshop concluded that: 1) the R0 likelihood profiles should be based on recruitment deviations that sum to zero and also examine how the recruitment deviations change with R0; and 2) the prediction skill of the model is an important model diagnostic, with the mean absolute scaled error (MASE) (Hyndman and Koehler 2006) being an important metric. The R0 likelihood profiling is used to examine the influence of each data component on the overall population scale and to assess whether the relative data weightings are appropriate and/or whether the model is mis-specified. (Lee et al. 2014).

In this paper, we re-examine the R0 likelihood profiles and estimate the prediction skill of the 2020 NPALB base case model through hindcasting. In the 2020 assessment, the R0 profiles were performed using recruitment deviations that did not sum to zero, and changes in the estimated recruitment deviations were not examined. There was no attempt to examine the prediction skill of the base case model in the 2020 assessment.

Methods

First, a new set of R0 likelihood profiles were performed using largely the same process as the 2020 assessment. The likelihood profile consisted of running a series of models with the ln(R0) parameter fixed at a range of values above and below that estimated by the base case model examining the likelihoods of the various data components. The only methodology difference between the R0 profiles from the 2020 assessment and this study was that the recruitment deviates for this study were forced to sum to zero by setting the 'do_recdev' switch in the SS control file to '1', instead of '2' in the 2020 assessment. In addition, we also plot the changes in the estimated recruitment deviations with respect to changes in the ln(R0) parameter, and compare the R0 profiles from this study and the 202 assessment.

Second, we examine the prediction skill of the 2020 base case model by hindcasting the model and calculating the MASE for the primary abundance index of the model. Following Kell et al. (2022), the MASE of a model with time 1 to *T* for a prediction horizon of *h* time steps and *n+1* predictions from replicate models with 0 to *n* time steps removed from the terminal end (i.e., a retrospective peel of *n* time steps), was calculated as:

$$MASE = \frac{\frac{1}{n+1}\sum_{t=T-n}^{T}|y_t - \hat{y}_{t|t-h}|}{\frac{1}{n+1+h}\sum_{t=T-n-h}^{T}|y_t - y_{t-h}|}$$

where, the mean absolute prediction error (i.e., the numerator) was calculated as the mean of the absolute differences between the observations *y* at time *t* (y_t) and the predictions of the observations made *h* steps previously ($\hat{y}_{t|t-h}$) for all the predictions made (*n+1*); and scaled by the mean absolute naive prediction error (i.e., the denominator) calculated as the mean of the absolute differences between the observations *y* at time *t* (y_t) and the naive predictions of the observations made *h* steps previously (y_{t-h}) for all the observations (*n+1+h*) between the first time step used (*T-n-h*) and time step *T* .

A series of SS models were developed from the 2020 base case model, with a retrospective peel of 0 to 5 years and a prediction horizon of 1 to 5 years. The prediction horizons of 2 and 5 years are of special importance to this study because the terminal years of data are typically 2 years prior to the year when the NPALB assessments are performed, and the NPALB assessments are currently conducted on a 3-year cycle. For example, the terminal year for the 2020 base case model was 2018, and the next NPALB assessment is expected to be conducted in 2023.

Results and Discussion

The likelihood profiles of log(R_0) from the 2020 assessment and the current study, where the estimated recruitment deviations summed to zero, provided largely consistent information about the overall population scale of the base case model but with fleet-specific differences in the relative importance of the size composition data (Figs. 1-3). The changes in the likelihood per unit change in log(R_0) were substantially higher when the estimated recruitment deviations summed to zero but it is important to note that the locations of the likelihood minima with respect to the F09 abundance index and the size compositions were relatively consistent for both sets of profiles. The location of the likelihood minima with respect to the F09 abundance index was slightly lower in the 2020 assessment [log(R_0) = 11.9] than when the recruitment deviations summed to zero [log(R_0) = 12.0] (Fig. 2). The location of the overall likelihood minima with respect to the size compositions was similar for both the 2020 assessment and when the recruitment deviations summed to zero [log(R_0) = 12.0] (Fig. 3). The shape of the likelihood profile from each fleet appeared to be similar but the relative importance of the size composition data from individual fleets to the overall likelihood profile was substantially different between the two profiles. For example, the Japanese pole-and-line fleet in Area 3 and Quarter 2 (F21) had the most informative size composition data in the 2020 assessment but the most informative size composition data for this study came from the Eastern Pacific Ocean Surface Fleet (F33).

Using recruitment deviations that summed to zero resulted in important changes to the pattern of recruitment deviations as the log(R_0) changed (Fig. 4). However, it is not clear cut that using recruitment deviations that summed to zero substantially improved the R0 profiles or simplified their interpretation. For example, the recruitment deviations in the R0 profiles from the 2020 assessment had a monotonically increasing trend with decreasing log(R_0), which was expected. However, when the recruitment deviations were assumed to sum to zero, the recruitment deviations for most years also appeared to increase trend with decreasing log(R_0). The increasing trend in recruitment deviations for most years were compensated by the very large decreases in recruitment deviations in the terminal years, when the model was relatively uninformative on the recruitment deviations. Our interpretation of these results are that R0 profiles remain useful for identifying consistency (or lack thereof) between data sources on the estimated population scale. However, the use of R0 profiles to identify the importance of specific fleets would need further work because the recruitment deviations can interact with different data sources in multiple ways, which can result in different scales of misfits.

The 2020 base case model appeared to exhibit some prediction skill over 1 - 2 years but not so over longer periods (Table 1). The MASE has a relatively simple interpretation, with a score of 0.5 indicating that the model has forecasts that are twice as accurate as naive forecasts. Therefore, the MASE scores indicated that the 2020 base case model had approximately 16% and 10% improvements in forecast accuracy of the F09 abundance index over horizons of 1 and 2 years, respectively, as compared to naive forecasts. However, the MASE for 3 or more years were >1, which indicated that forecasts of 3 or more years did not exhibit any prediction skill over naive forecasts. Given these results, the ALBWG can be confident that the 2020 base case model has sufficient skill to estimate and forecast current stock status with a horizon of 2 years. However, it may be useful to update the F09 abundance index during the intervening years to monitor the relative status of the spawning stock biomass.

References

ALBWG. 2020. Stock assessment of albacore tuna in the North Pacific Ocean in 2020. Page 109. International Scientific Committee for Tuna and Tuna-like Species in the North Pacific Ocean,

Stock Assessment.
Carvalho, F., H. Winker, D. Courtney, M. Kapur, L. Kell, M. Cardinale, M. Schirripa, T. Kitakado, D. Yemane, K. R. Piner, M. N. Maunder, I. Taylor, C. R. Wetzel, K. Doering, K. F. Johnson, and R. D. Methot. 2021. A cookbook for using model diagnostics in integrated stock assessments. Fisheries Research 240:105959.
Hyndman, R. J., and A. B. Koehler. 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22(4):679–688.
Kell, L. T., R. Sharma, T. Kitakado, H. Winker, I. Mosqueira, M. Cardinale, and D. Fu. 2021. Validation of stock assessment methods: is it me or my model talking? ICES Journal of Marine Science 78(6):2244–2255.
Lee, H. H., K. R. Piner, R. D. Methot, and M. N. Maunder. 2014. Use of likelihood profiling over a global scaling parameter to structure the population dynamics model: AN example using blue marlin in the Pacific Ocean. Fisheries Research 158:138–146.
Maunder, M., and K. Piner. 2015. Contemporary fisheries stock assessment: many issues still remain. ICES Journal of Marine Science 72(1):7–18.
Methot, R. D., and C. R. Wetzel. 2013. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Fisheries Research 142:86–99.
Mohn, R. 1999. The retrospective problem in sequential population analysis: An investigation using cod fishery and simulated data. ICES Journal of Marine Science 56(4):473–488.

Table 1. Prediction skill of the 2020 North Pacific albacore base case model on the F09 abundance index over a horizon of 1 to 5 years and a retrospective peel of 0 to 5 years, using mean absolute scaled error (MASE) as the metric of skill.

Horizon (years)	Mean absolute prediction error (index units)	Mean absolute naive prediction error (index units)	Mean absolute naive prediction error for entire time series (index units)	Mean absolute scaled error (MASE)
1	5.86	6.80	6.78	0.86
2	6.50	7.12	7.33	0.91
3	7.21	6.80	6.40	1.06
4	8.25	5.56	7.73	1.48
5	9.93	4.67	7.35	2.13

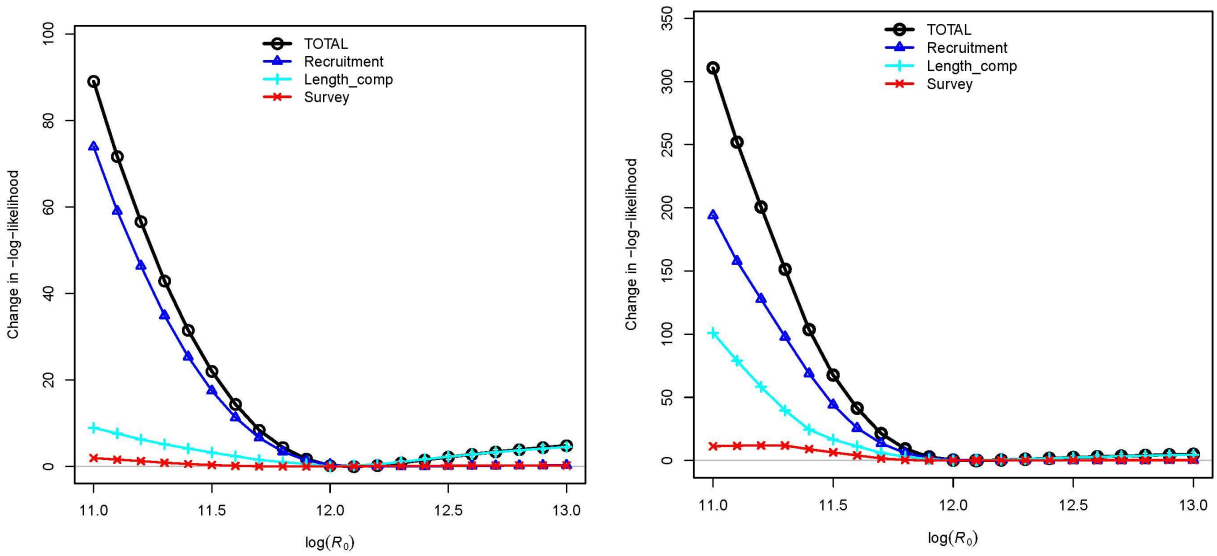

Figure 1. Likelihood profiles of virgin recruitment [log(R_0)] with respect to the main data components for the 2020 assessment (left) and the current study (right).

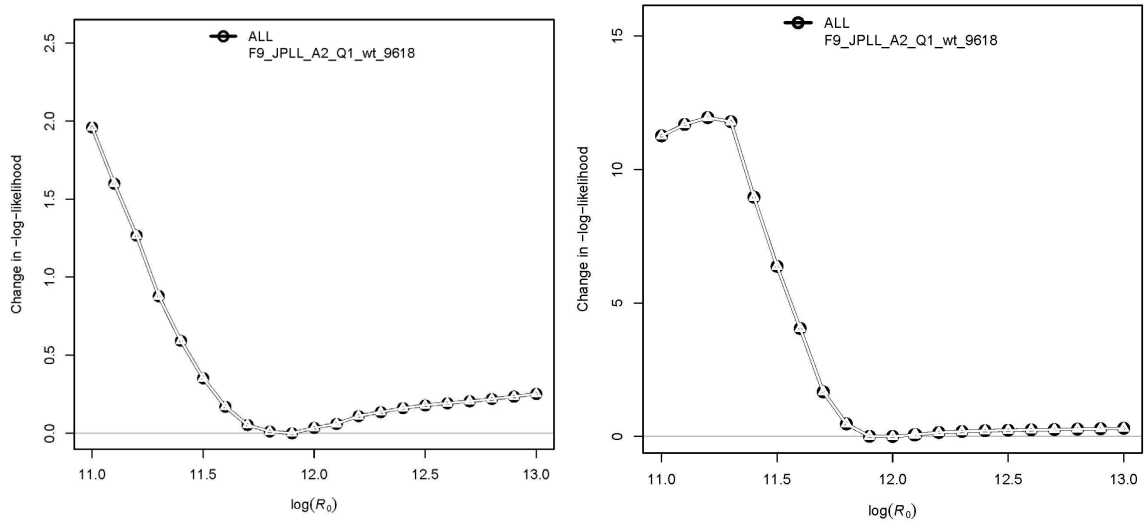

Figure 2. Likelihood profiles of virgin recruitment [log(R_0)] with respect to the F09 abundance index for the 2020 assessment (left) and the current study (right).

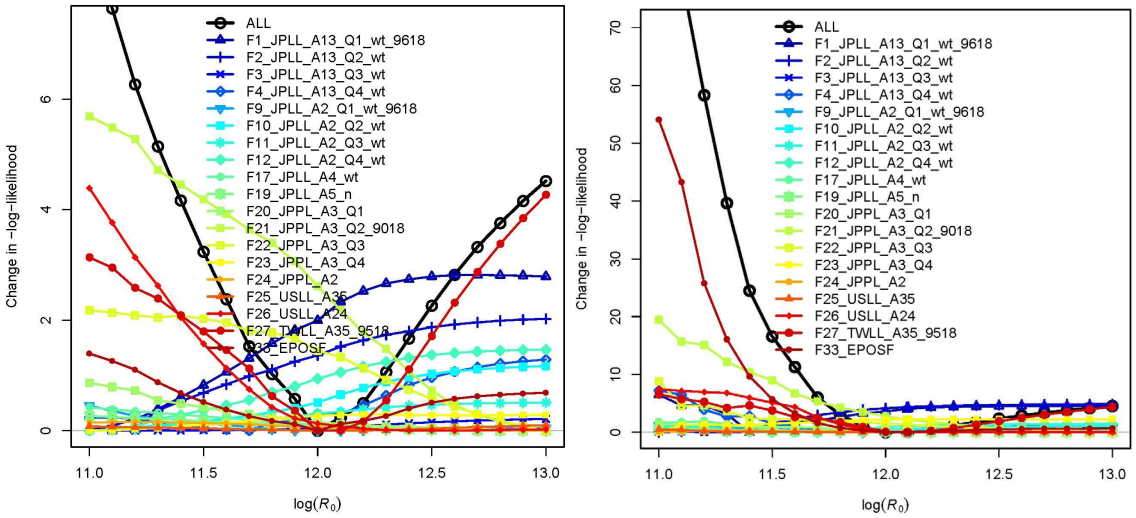

Figure 3. Likelihood profiles of virgin recruitment [log(R_0)] with respect to the size compositions for the 2020 assessment (left) and the current study (right).

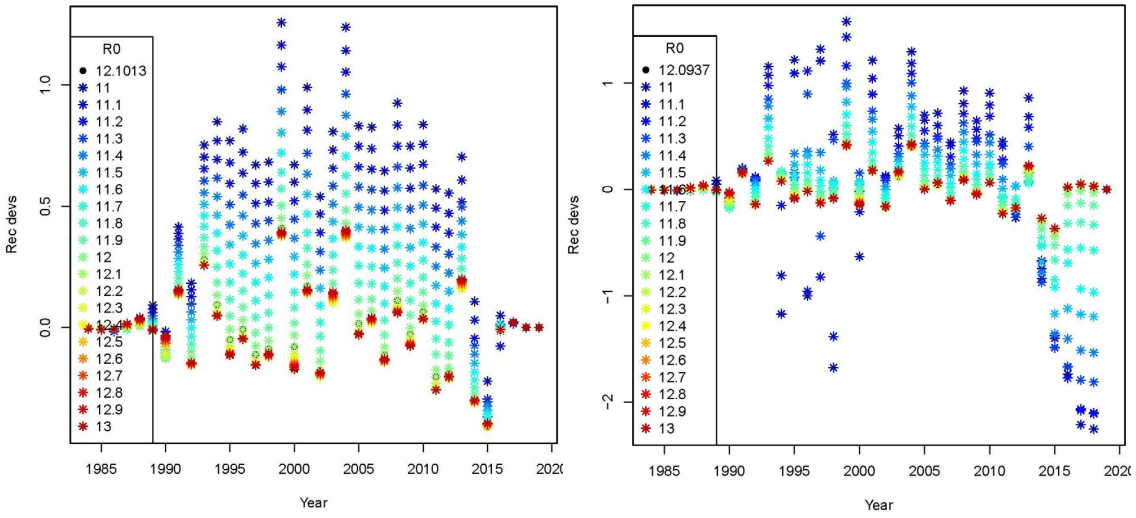

Figure 4. Estimated recruitment deviates from the virgin recruitment [log(R_0)] profiles of the 2020 assessment (left) and the current study (right).