**INTER-AMERICAN TROPICAL TUNA COMMISSION**

**SCIENTIFIC ADVISORY COMMITTEE**

**11TH MEETING**

**San Diego, California (USA)**
**11-15 May 2020**[1]

# DOCUMENT SAC-11 INF-F REV

## IMPLEMENTING REFERENCE POINT-BASED FISHERY HARVEST CONTROL RULES WITHIN A PROBABILISTIC FRAMEWORK THAT CONSIDERS MULTIPLE HYPOTHESES

Mark N. Maunder, Haikun Xu, Cleridy E. Lennert-Cody, Juan L. Valero, Alexandre Aires-da-Silva, Carolina Minte-Vera

**CONTENTS**



**ABSTRACT**

A risk assessment is developed to implement reference point-based fishery harvest control rules within a probabilistic framework that considers multiple hypotheses. This pragmatic approach is a compromise between computational demands, complexity, and statistical rigor. It acknowledges the need to weight models based on information in the available data, but does so in a context where the complexity of fisheries stock assessment models prevents strict adherence to statistical rigor. The main features of this approach are: 1) hypotheses about states of nature are represented by alternative stock assessment models with specific model structure, data use and parameters; 2) hypotheses are grouped into a hierarchical framework, which highlights similarities among models thereby avoiding that any one hypothesis, or overarching hypothesis, inadvertently dominates the outcome of the risk analysis, and

[1] Postponed until a later date to be determined

facilitates model development and weight assignment; 3) sub-hypotheses represent models with parameters that cannot be reliably estimated within the assessment model and are therefore fixed in the models; 4) multiple metrics are used to evaluate the reliability of the models and the plausibility of the hypotheses they represent; 5) model fit only plays a limited role in metrics used to evaluate models; 6) an efficient approach to eliminate unlikely hypotheses. This approach was illustrated by applying it to the stock assessment of EPO bigeye tuna and was used to evaluate a) probability statements about the current status relative to reference points and b) probability statements about the fishing mortality relative to reference points under different management scenarios.

## 1. INTRODUCTION

The management advice formulated by the IATTC scientific staff, on which the Commission bases its decisions for conserving and managing the tuna stocks in the eastern Pacific Ocean (EPO), has traditionally been based on the "best assessment" approach. This involves making a series of assumptions about the stocks, their biology, the fisheries and the environment, then fitting a stock assessment model to the available data to estimate the model parameters and quantities of interest (*e.g.* $F_{MSY}$, the fishing mortality that corresponds to the maximum sustainable yield for that stock). This single model, called the "base case", has hitherto formed the basis for the staff's management advice. Although confidence intervals are usually also presented for the quantities of interest, along with sensitivity analyses (comparisons with other models), the management advice does not fully account for uncertainties in the estimates, nor does it address alternative management decisions and their potential implications. The staff has therefore been investigating and applying two related methodologies that overcome these limitations: management strategy evaluation (MSE) and risk analysis.

**Risk analysis**, of which there are many types, takes uncertainty into account quantitatively in its management advice. It has a long history in fisheries: for example, decision tables that predict the outcomes of a range of management actions under different states of nature, and the probabilities associated with those states of nature, have been used for decision-making in several fisheries (Punt and Hilborn 1997). More recently, probability statements have been integrated directly into **harvest control rules** (HCRs), under which specified combinations of stock and fisheries status and trends trigger predefined management actions, and explicit estimates of uncertainty are used to evaluate the probability statements. HCRs are often based on target, limit, and threshold reference points: for example, the IATTC's HCR for tropical tunas establishes that "*if the probability that* F *will exceed the limit reference point* ($F_{LIMIT}$) *is greater than 10%, as soon as is practical management measures shall be established that have a probability of at least 50% of reducing F to the target level* ($F_{MSY}$) *or less, and a probability of less than 10% that F will exceed* $F_{LIMIT}$."(Resolution C-16-02).

**MSE** has become the gold standard for addressing uncertainty (Butterworth 2017, Punt *et al*. 2016), and has been widely used both nationally and internationally, including by all five regional fisheries management organizations for tuna (t-RFMOs: IATTC, IOTC, WCPFC, ICCAT, CCSBT) (Nakatsuka *et al*. 2017). A management strategy is the combination of data, data analyses, and HCRs to determine what management measures are to be taken to meet a set of management objectives, and MSE compares the performance of different management strategies in meeting those objectives. In MSE, computer simulations of population and fishery dynamics are used to calculate relevant simulated fisheries performance metrics, such as annual variability in seasonal fishery closure length, under different types of uncertainty, including alternative states of nature, which are represented by different operating models in the simulation analysis.

The methods for developing these operating models are similar to those used to incorporate uncertainty into a risk analysis. Because implementing an MSE is a long and complex process, the transition from a "best assessment" approach typically occurs in stages, and a risk analysis can serve as an intermediate

step in this process.

To conduct a risk analysis in an HCR framework, uncertainty is expressed as alternative *states of nature*, such as the state of the stock (*e.g.* biomass (*B*)), the factors that control the dynamics of the stock (*e.g.* natural mortality (*M*), selectivity, stock-recruitment relationship), or how the data are related to the stock (*e.g.* catchability, sampling distributions). This requires implementing models that represent the alternative states of nature and using those models to calculate the probability of each state of nature being true.

Three categories of uncertainty are addressed in risk analysis: 1) *parameter uncertainty*, 2) *model structure uncertainty*, and 3) *uncertainty about the future* (*e.g.* process variation). This document considers only the first two; the third will be addressed as the methodology is further developed. There is ambiguity in the difference between parameter uncertainty and model structure uncertainty, but they can both be used to represent alternative states of nature.

One of the critical steps in conducting a risk analysis is to obtain estimates of parameter and model structure uncertainty. While estimating parameter uncertainty is straightforward, and common in stock assessment, estimating model structure uncertainty is more difficult. The model structure uncertainty used in a risk analysis should be informed by three information types: expert knowledge, theory, and available data. The goal of a risk analysis to evaluate a probabilistic HCR is to use these three types of information to develop probability distributions that represent the uncertainty in the estimates of the current status of the stock (*e.g.*, biomass, fishing mortality) relative to the associated target and limit reference points. Since quantities of interest are typically the ratios of the current status to a reference point (*e.g.* $F_{cur}/F_{MSY}$), probability distributions are calculated for the ratios, rather than for their individual components.

Various methods (*e.g.* Privitera-Johnson and Punt 2020) can be used to calculate probability distributions for quantities of interest, most of which use approximations due to the computationally demanding nature of contemporary statistical integrated stock assessment models. These methods are not reviewed or discussed in detail in this study; instead, we describe a somewhat idealized approach, and then use this to develop a practical approach more suitable for the management of many species (among them tropical tunas in the EPO). Approximations (see Sections 3-4) are used to create the probability distributions for the quantities of interested and these are evaluated by comparison to Bayesian posteriors.

As in any risk analysis, the models that represent the alternative states of nature must be assigned weights. At extremes of the spectrum of possible weighting schemes are three options, relative weighting based solely on model fit, equal weighting of all models, and expert opinion. Contemporary fishery stock assessments are complex, and using model fits alone to weight the various models may not be appropriate, or even possible, because biologically implausible models may be heavily weighted simply because they fit the available data better. Equal weighting of all models is a commonly used alternative, but may result in biased advice, as, for example, when many models representing similar states of nature result in a higher effective weight than a single model representing a different state of nature. A common use of expert opinion is the selection of a single base case model to use for providing management advice as selected by the assessment author. However, typically the choice is a somewhat subjective ad hoc approach of model development and testing that uses both model fit and expert opinion to choose the model. A more organized systematic approach is needed. To avoid the problems associated with these three options, we propose a range of metrics related to the reliability of a model, in addition to model fit, and combine them to assign a probability to each model. We also use a hierarchical framework to present possible states of nature to facilitate model development, as illustrated in the following application to bigeye tuna in the EPO.

### 1.1. Application to bigeye tuna

The stock assessments for bigeye tuna in the EPO have become problematic in recent years, for various reasons (SAC-09 INF-B). In particular, an increase in the estimated recruitment starting in the mid-1990s results in an apparent "two-regime" pattern in recruitment, with estimates about doubling after 1993. This "recruitment shift" coincides with the rapid geographic expansion of the purse-seine fishery on fish-aggregating devices (FADs) in the equatorial EPO in the mid-1990s, and the sudden and dramatic increase in associated catch. Some of the hypotheses proposed to explain this shift (Aires-da-Silva *et al*. 2010, Valero *et al*. 2019) ascribe the increase to a modelling artifact, while others postulate that recruitment really did increase (see Valero *et al*. (2019) and Punt *et al*. (2019) for details). In addition, the assessment results have become highly sensitive to new data points in the indices of relative abundance derived from the longline fishery (SAC-09 INF-B), due perhaps partially to the spatial contraction of that fishery. These and other issues, such as systematically poor model fits to length-composition data and the possibility of population structure not being captured in the model, have yet to be resolved, and some may never be completely resolved. Therefore, management advice could be improved by incorporating uncertainty when evaluating the IATTC HCR via a risk analysis.

The goal of the risk analysis for tropical tunas in the EPO is to determine the probability of exceeding the target and limit reference points under alternative management actions, such as fishery closures. The IATTC HCR for tropical tunas (Resolution C-16-02, paragraph 3) states:

a. "The scientific recommendations for establishing management measures in the fisheries for tropical tunas, such as closures, which can be established for multiple years, shall attempt to prevent the fishing mortality rate ($F$) from exceeding the best estimate of the rate corresponding to the maximum sustainable yield ($F_{MSY}$) for the species that requires the strictest management.
b. If the probability that $F$ will exceed the limit reference point ($F_{LIMIT}$) is greater than 10%, as soon as is practical management measures shall be established that have a probability of at least 50% of reducing $F$ to the target level ($F_{MSY}$) or less, and a probability of less than 10% that $F$ will exceed $F_{LIMIT}$.
(…)
c. If the probability that the spawning biomass ($S$) is below the limit reference point ($S_{LIMIT}$) is greater than 10%, as soon as is practical management measures shall be established that have a probability of at least 50% of restoring $S$ to the target level (dynamic $S_{MSY}$) or greater, and a probability of less than 10% that $S$ will descend to below $S_{LIMIT}$ in a period of two generations of the stock or five years, whichever is greater."

Various estimates are needed to evaluate the IATTC HCR in a risk analysis framework, including the probability of exceeding the target and limit reference points, both currently and in the future. Ideally, the estimated probabilities would encompass uncertainty in both current and future conditions but, as noted above, this study focuses on the calculation of uncertainty for the current status of bigeye tuna, as represented by the stock assessment estimates with respect to reference points, and leaves the future conditions for further work. The IATTC has started a MSE process for tropical tunas (IATTC MSE Workplan) that will define candidate reference points, HCRs, performance measures to be tested along with timeframes for specific objectives of alternative management strategies (Valero and Aires-da-Silva 2020).

## 2. METHODS

We first describe an "idealized" approach and then describe a more practical approach that we propose, illustrating its use with an application to bigeye tuna in the EPO. In what follows, we use the term

hypothesis to refer to a state of nature. Each hypothesis is associated with one or more models that represent the hypothesis in a stock assessment context.

### 2.1. "Idealized" approach

In the "idealized" approach, all available data, alternative models, population dynamics theory, and expert opinion are used to calculate the probability distributions of the current status of the stock exceeding the target and limit reference points. In the case of bigeye in the EPO, such an approach is desirable because several alternative models (see 'Reference models' in Xu *et al*. 2020) have been formulated that explain the observed fishery data for bigeye, but there is uncertainty about which one best captures the reality of the population dynamics and fishery processes. The idealized approach takes into account any external data sources not used in the assessment, population dynamics theory, and expert opinion to develop the structure of the alternative models, priors for their parameter values, and then uses the data within each model to update this information. This, in conjunction with integrating over the parameters to provide posterior distributions for the quantities of interest, is the foundation of the Bayesian approach proposed by Punt and Hilborn (1997). Thus, the idealized approach is to use a Bayesian analysis integrating all the relevant data directly into the model or through data-based priors, parameterizing alternative model structure as much as possible, and estimating all the model parameters. Posterior distributions for the quantities of interest can then be used for management advice and for evaluating reference points in a probabilistic framework as required by some HCRs.

In practice, however, implementation of the idealized approach is problematic for several reasons. First, using standard statistical methods of assessing model reliability, such as AIC, do not work well for complex, highly parameterized models fit to lots of data of different types. Furthermore, in fisheries, model assumptions are typically violated, invalidating the use of such likelihood-based measures. In addition, data weighting for fisheries models is usually not as well defined as in simple statistical models. Therefore, other metrics of model reliability need to be evaluated.

A second issue is that not all hypotheses about uncertainty in model structure can be simply represented by a set of parameters. There are algorithms to estimate the probability of the model from the data (*e.g.* reverse jump MCMC), but these are seldom used in fisheries stock assessment. A more practical approach is to parameterize the model structural uncertainty and allow the parameter estimation to represent the uncertainty in the model structure. For example, the additional parameter in the Deriso (1980) stock recruitment model can be used to represent the two most common stock-recruitment models, Beverton-Holt and Ricker, and all the models in between. However, this is not always possible or practical (*e.g.* when evaluating alternative fishery structures). One specific case is when there is the possibility that a data set is not representative, and alternative models are run with and without that data set. Conceptually, the inclusion of a data set can be parameterized by estimating the standard deviation of the likelihood that measures the fit to the data. However, this just produces a model that is a compromise among the data sets and does not reflect that one data set could be unrepresentative and completely misleading (Schnute and Hilborn 1993).

A further issue is that in fisheries there is often a lack of appropriate data to estimate the different parameters. Preferably, the data already in the model would inform all choices of the model structure and the parameter values, but in practice, particularly for stock assessment applications, this is not the case due to lack of information. There is also the danger when information is insufficient that non data-based priors, even those that were chosen to be "uninformative", can influence the results. Thus, some assumptions based on expert judgement need to be made regarding parameters.

Finally, Bayesian models are typically fit using Markov chain Monte Carlo (MCMC) methods, but this approach to model fitting can be prohibitively slow, a particular disadvantage when multiple models need

to be evaluated. Although MCMC would be the preferred methodology for estimating probability distributions for quantities of interest (*e.g.*, $F_{cur}/F_{MSY}$), a simplification, such as using normal approximations based on frequentist statistics for possibly non-normal probability distribution functions, is necessary. In addition, the possible combinations of models can be very large and an approach is needed to limit the number of models considered.

## 2.2. Practical approach

There are 5 main steps involved in the implementation of the practical approach: 1) establishing a hierarchy of hypotheses and models; 2) defining a weighting system for hypotheses and models; 3) calculating the probability distributions for quantities of interest for a model; 4) combining probability distributions across models; 5) presenting the results in the form of a risk analysis. These steps are explained below.

### 2.2.1. 1) Establishing a hierarchy of hypotheses and models

In a risk analysis, hypotheses about the state of nature are represented by models. Typically, there will be multiple models that could be used to represent a single hypothesis. In addition, hypotheses may be nested, adding further complexity. In reality, there is an unlimited combination of hypotheses, and possible models to represent them, and some prudence is needed when defining the models and combinations. Therefore, to facilitate the development of models that represent hypotheses about the state of nature, we use a flow chart to represent a hierarchical structure of hypotheses that has three levels: 1) Overarching hypotheses; 2) hypotheses; 3) sub-hypotheses. These levels have different functions and are described below.

#### 2.2.1.A Level 1: Overarching hypotheses

The overarching hypotheses correspond to broad states of nature (e.g. the number of stocks). They are represented by a variety of models of different complexity and that may use different data. These overarching hypotheses are put in the first level because it is difficult, impossible, or inappropriate to evaluate them simply by the fit of models to data. Therefore, the overarching hypotheses rely solely on expert opinion to provide weights and the weighting of any models under an overarching hypothesis are conditional on those weights.

#### 2.2.1.B Level 2: Hypotheses

Level 2 hypotheses are specific hypotheses that can be represented by a model and can be differentiated by fitting the models to data. Level 2 may be divided into sub-levels (A, B, …) where each sub-level addresses an issue in the assessment. Models are developed for each of these sub-levels and typically need to be used in combination to solve all the assessment issues. The sub-levels provide a convenient way to organize the models and may aid in assigning weights (see Section 2).

#### 2.2.1.C Level 3: SUB-HYPOTHESES

Level 3 sub-hypotheses are evaluated differently than the Level 2 sub-levels to avoid the influence of data, reduce the number of analyses, or for convenience. Level 3 sub-hypotheses are typically encompassed by a single hypothesis, can be represented by restricting a model (*e.g.* fixing the value of a parameter, such as steepness), and are applied to most, if not all, models on Level 2.

### 2.2.2. 2) Defining a weighting system for hypotheses and models

Once a hierarchy of hypotheses has been established (Step 1), various sources of information (both internal and external to each model) can be used to construct a system to evaluate which hypotheses are considered more likely than others. This is a necessary step because assigning the same weight (reliability)

to all hypotheses could introduce biases into the management advice if some hypotheses are, in fact, highly unlikely. Weights need to be assigned to each of the models (representing hypotheses) in the hierarchy outlined in Section 1 (mainly at Level 2). The model weights must then be rescaled so that they represent probabilities and can be used in the risk analysis. This is achieved through a weighting system involving the following steps: a) establish weight categories that will be used to assign weights to models and hypotheses; b) select weight metrics to be used to weight models and hypotheses; c) assign weights to models and hypotheses, and rescale weights so they can be used in a probabilistic framework; and d) ensure the number of hypotheses is practical (as too many or too similar hypotheses may be difficult or impractical to evaluate). Each of these steps is discussed in detail below.

#### 2.2.2.a 2a) Weight categories

To assign weights to the various models and hypotheses, it is preferable to establish a system of discrete weight categories. The complex and uncertain nature of fisheries stock assessment models makes assigning continuous quantitative values as weights representing model structure uncertainty difficult, and as a result, assigning the weights is often qualitative and subjective. An attempt is made to facilitate and standardize the weighting by creating discrete weight categories to try to capture the degree of reliability of each model. The numerical values assigned to each weight category are themselves subjective. We suggest the following categories and numerical weights be used for all weight metrics except the metric for overall model fit (*W(Fit)*; see Section 2.2.2b).

| Weight Category | Value |
|---|---|
| None: | 0 |
| Low: | 0.25 |
| Medium: | 0.5 |
| High: | 1.0 |

#### 2.2.2.b 2b) Weight metrics

Using solely model fit to weight alternative models and hypotheses may not be possible, or even appropriate. Therefore, we develop a set of metrics related to the reliability of a model, in addition to model fit, which will later be combined to assign weights to each model representing a hypothesis. These metrics include:

a. *W(Expert):* Expert opinion, assigned "a-priori", without consideration of model fit.
b. *W(Convergence):* Model convergence criteria of the estimation algorithm.
c. *W(Fit):* The fit of the model to the data.
d. *W(Plausible parameters):* The plausibility of the estimates of the parameters representing the hypothesis.
e. *W(Plausible results):* The plausibility of the model results.
f. *W(Diagnostics):* Reliability of the model based on diagnostics.

The weight given to each model is then the product of the individual component weights, once each of those components has been rescaled (See Section 2c):

$$W(model) = W(Expert) \times W(Convergence) \times W(Fit) \times W(Plausible\ parameters) \times W(Plausible\ results) \times W(Diagnostics)$$ **[Equation 1]**

Depending on the application (*i.e.* particular stock the weighting system may be applied to), additional weighting metrics should be considered to represent how well the models address the issues represented in Level 2 of the hypothesis hierarchy.

The *W(Diagnostics)* component is calculated based on a variety of diagnostics. The same set of weight categories as presented above (*i.e. None*, *Low*, *Medium*, *High*) are used for calculating each of the diagnostic weights. The weights of each set of diagnostics are added together (rather than multiplied) to ensure that the individual diagnostics are not overweighted in the calculation of *W(Model)*:

*W(Diagnostics) = W(ASPM, $R_0$, Catch curve) + W(Retrospective analysis) + W(Composition residuals) + W(Index residuals) + W(Recruitment residuals)* **[Equation 2]**

Each metric is described below.

**a.** *W(Expert):* represents the subjective judgment of the experts, taking into consideration past experience with the specific assessment, other assessments of that species, assessments of other species, and other biological, ecological and related factors. Population dynamics theory is also taken into consideration.

**b.** *W(Convergence):* represents how well the parameter estimation procedure worked. If the Hessian matrix is not positive definite, it is considered that the model has not converged and *W(Convergence)* is given zero weight, independent of the other criteria.

If the estimation procedure has not converged correctly, the parameter estimates may not be correct and therefore the model does not represent the hypothesis well. It also may indicate that the model is a misspecified representation of the hypothesis or that the hypothesis is not consistent with the data. However, effort should be made to understand the problem and get a positive definite Hessian and low maximum gradients before assigning weights (*e.g.*, by checking parameter values relative to their estimating bounds, trying different starting parameter values, changing the phases in which each parameter is estimated, increasing the maximum number of function evaluations in the function minimizer). Jittering starting values and phases can also be used to investigate the stability of model results and determine whether local minima are present. Models with many local minima that produce different management results may need to be assigned lower weights, particularly if this problem will not otherwise be represented in the probability distributions of the quantity of interest for that model. The maximum gradient component or other aspects of model convergence could also be investigated as possible factors to use for determining this weighting metric.

**c.** *W(Fit):* represents how well the model fits the data. This measure is based on the overall fit to the data and is not related to the residual pattern, which is covered by *W(Diagnostics),* as outlined below. The total negative log-likelihood could be used to measure overall fit. However, models representing different hypotheses often have different numbers of parameters (or data, described below) and adding parameters typically improves the fit, which is not accounted for by the total negative log-likelihood. Therefore, AIC or similar statistics should be used to adjust for the number of estimated parameters. Burnham and Andersen (1998) provide the following guidelines regarding the use of AIC to evaluate a collection of models relative to the "best" model, where the best model is the model with the lowest AIC. The larger difference in AIC from the best model (Δ AIC) the stronger the evidence against that model (i.e. the lower the Δ AIC (or equivalently the AIC) the more support for that model):

$\Delta AIC \leq 2$ no evidence against the model
$2 < \Delta AIC \leq 4$ weak evidence against the model
$4 < \Delta AIC \leq 7$ definite evidence against the model
$10 < \Delta AIC$ very strong evidence against the model

where $\Delta AIC = AIC - AIC_{min}$, and $AIC_{min}$ is the lowest AIC over all models.

To address the complexities of assessing overall model fit in stock assessment models, we apply more liberal criteria than the guidelines shown above to ensure that all models are considered, but we realize that our criteria are inconsistent with those guidelines and also with the probability distributions estimated from the model for the management quantities. Our procedure gives *High* weight to the model with the lowest AIC (best fit) and *Low* weight to the model with the highest AIC (worst fit) and a linear function of AIC in between:

$$W(Fit) = Low + (High - Low) \times (1 - [\Delta \text{ AIC} / \max(\Delta \text{ AIC})]) \quad \textbf{[Equation 3]}$$

However, the calculation of AIC gets complicated when the models do not use the same data, cover a different time frame, or use different data weighting. AIC is not valid if the data differs among models and applying the AIC based only on the data common among models favors the model with less data which is expected to fit the common data better because it is less constrained and is likely to be the minimum AIC model. We suggest three possible ways to deal with this issue:

1. Calculate the AIC for models that use the same data and rescale the weights among these models to sum to one.
2. For models with an additional data set that is specific to an additional parameter that is being estimated in the model (*e.g.* age-length data and estimating growth), calculate the AIC for all models without this data.
3. When calculating the weight for models with less or downweighted data, replace Δ AIC in the numerator of the formula above with the quantity ($\max(AIC_{limited})-AIC_{limited}$), where $AIC_{limited}$ refers to the AIC calculated using only the data common to all models and are not down-weighted:

$$W(Fit) = Low + (High - Low) \times \text{Max}[1, (1 - (\max(AIC_{limited})-AIC_{limited}) / \max(\Delta AIC))] \quad \textbf{[Equation 4]}$$

**d.** *W(Plausible parameters):* represents the realism of the estimates of the additional parameters that are added to represent the hypothesis. In many cases, a parameter that is added to the model to solve the perceived cause of the model misspecification is instead used by the model to compensate for the actual, but different, model misspecification. This may be apparent in estimates of that parameter that are unrealistic. This could be judged based on data, but if those data were in fact available, then they should be included in the model directly or used to create a prior for the parameter. Also, each set of parameters representing a hypothesis is different for each model. Therefore, evaluation of this metric is somewhat subjective. It should be noted that even though an unrealistic value is estimated, a more realistic value may also be supported by the data and so the uncertainty in the parameter estimate should also be considered.

**e.** *W(Plausible results):* this metric is very subjective and care needs to be taken that the final outcome is not controlled by this metric. Stock assessments produce many results, so providing guidance on evaluation of this metric is difficult. However, one factor to consider is unrealistically high or low estimated fishing mortality.

**f.** *W(Diagnostics):* represents the reliability of the model. *W(Diagnostic)* serves two purposes: 1) to evaluate if the data support the hypothesis represented by the model, assuming the model is correctly specified, and 2) evaluate if the model correctly specifies the hypothesis. As regards the first purpose, a model may fit the data well, as measured by an objective function (*e.g.* the likelihood) and adjusting for the number of parameters estimated (*e.g.* AIC), but if that model violates assumptions or provides unrealistic results then it suggests that the hypothesis may not be correct. As regards the second purpose, a model may violate assumptions or provide unrealistic results because it incorrectly specifies the hypothesis. In general, a model that violates assumptions or provides unrealistic results

should not be used to represent that hypothesis, and, the model should not be used in the risk analysis. However, it is still desirable to represent the hypothesis in the risk analysis, and a model that better represents the hypothesis should be found, if possible. In some cases, it may be difficult to determine if the violations of assumptions are due to the hypothesis being incorrect or the model incorrectly specifying the hypothesis.

Here we outline how the model diagnostics are used to define weights for each component of *W(Diagnostics)*.

**W(ASPM, $R_0$, Catch curve)**

The $R_0$ likelihood component profile, the Age-structured production model (ASPM) diagnostic, and the catch curve diagnostic all provide information about the abundance information content of different data sets. They provide information about how influential the composition data are on the estimates of absolute abundance. Therefore, they are grouped together in one category for assigning the weights. The ASPM and catch curve diagnostics also show the influence of composition data on the estimated trends in abundance. Preferably, all data sets provide consistent information, or at least the indices of abundance provide most of the information on absolute abundance and trends in abundance. It should be noted that annual estimates of relative recruitment might be needed to extract information on absolute abundance from the indices and these recruitments are informed by the composition data.

Further information on these diagnostics is provided below. An overview of the algorithm to assign weights based on the $R_0$ profile diagnostic and the ASPM diagnostic is presented in Figure 1. This weight is then multiplied by the weight from the Catch Curve diagnostic:

$$W(ASPM, R_0, Catch\ curve) = W(ASPM, R_0) \times W(Catch\ curve)$$ **[Equation 5]**

*$R_0$ profile*

The $R_0$ likelihood component profile determines the information in the different data sets about absolute abundance and whether they are in conflict. Preferably, all the data provide the same information about absolute abundance. Any differences indicate possible model misspecification. In general, if there are differences, it is preferable that absolute abundance information come from the index of abundance and not from the composition data, particularly not from length composition data, assuming the index of abundance is considered representative. Achieving this may require down-weighting the composition data. However, the model will still be misspecified and the misspecification may be impacting results. Alternatively, the problem may be that the index is not be representative.

The following choice of weight categories is recommended: *High* weight should be given to models were the data provide consistent information, *Medium* weight should be given to models with inconsistent information, but the index has more influence, and *Low* weight should be given to models with inconsistent information and the composition data have more influence. However, the consistency of information may be difficult to determine. For example, one data source may provide information that biomass must be higher than a certain amount but provides no information about how high the biomass should be. This type of ramping of the likelihood profile may simply be a consequence of the type of data but may look like data conflict. Also, one data set may be uninformative and so not show conflict.

*ASPM and ASPM-Rdev Diagnostics*

The ASPM diagnostic is used to determine whether the impact of the catch on the index of abundance is driving the estimates of absolute abundance and/or is consistent with the information from the composition data, and whether information on recruitment variation is needed to extract the absolute

abundance information (using ASPM-Rdev, the ASPM diagnostic with the recruitment deviates estimated, and its confidence intervals). Preferably, for the fully integrated model that fits to all the data (*e.g.* index of abundance and composition data), ASPM, and ASPM-Rdev provide similar estimates of absolute abundance and trends over time. However, for stocks that show variable autocorrelated recruitment, and/or for which recruitment makes up a substantial component of the abundance, the ASPM is expected to perform poorly (*e.g.* estimate much larger absolute abundance) and recruitment deviates are needed. If variable autocorrelated recruitment is not evident (*i.e.* is not estimated by the fully integrated model) then differences in the abundance estimates between the fully integrated model and the ASPM indicates that either the composition data is driving the estimates, which can be confirmed by the $R_0$ profile, that the model is misspecified, or both. Large confidence intervals on biomass estimated by the ASPM-Rdev indicate that information on recruitment deviates (*e.g.* from composition data) may be needed to interpret the information on absolute abundance obtained from the impact of catch on the index of abundance. If the ASPM-Rdev confidence intervals are reasonable, then differences in the abundance estimates between the fully integrated model and the ASPM-Rdev indicates that either the composition data is driving the estimates, which can be confirmed by the $R_0$ profile, that the model is misspecified, or both.

*Catch Curve Diagnostic*

The catch curve diagnostic simply estimates the model parameters without fitting to the indices of abundance by only fitting to the composition data. Differences in estimates of abundance, both in absolute terms and in trends over time, between the catch curve diagnostic and the fully integrated model, indicate model misspecification. Differences in absolute abundance indicate that growth (particularly the asymptotic average length) may be misspecified, if fitting to length composition data, or that the dome-shaped form of the selectivity may be misspecified. Differences in trends indicates that growth or selectivity may have changed over time, but are modelled as time invariant. The weight given to each model should be based on the size of the differences. There may be overlap in the information provided between this diagnostic and the $R_0$ profile and the ASPM diagnostic.

*Retrospective analysis*

Retrospective analysis determines if the results change as additional years of data are added. Changes in estimates of biomass, recruitment, or fishing mortality that are large and are in the same direction as more years of data are added indicates model misspecification. The weight category assigned to *W(Retrospective analysis)* should be based on the magnitude of the difference and whether the change is always in the same direction. More quantitative measures can be used (*e.g.* Mohn's Rho), but we use a subjective approach.

**Residual analysis**

Model misspecification can be identified by characteristics of the residuals that are not consistent with the assumptions implicit in the likelihood functions used to fit the data. The misspecification could be in the system model (the population dynamics), the observation models that relate the system model to the data, or in the sampling distribution assumptions (the likelihood functions). Violation of model assumptions may manifest in the residuals in several ways. Here we focus on three characteristics: the magnitude of the residuals, trends in residuals (violating the independence assumptions), and the shape of the distribution of the residuals, including outliers, although we note that the distributional shape and outliers are typically indicative of lower-order issues and are therefore generally ignored.

When assigning weights, there are two important considerations. First, automatic data weighting (*e.g.* Francis weighting for composition data) is becoming the standard approach to fit data and therefore the

absolute magnitude of the residuals is not useful as a diagnostic when using automatic weighting. Second, trends in residuals are common and therefore expected, so it may be best to only consider substantial trends in residuals when determining weights. There may be quantitative criteria that would be appropriate, and these should be investigated.

Of course, preferably the model misspecification is corrected before the model is included in the risk assessment. The characteristics of the residuals may indicate what part of the model is misspecified. For example, if the absolute magnitude of the residuals is too large, this issue is often mitigated by estimating the quantity related to the variance of the likelihood function (*e.g.* the standard deviation of a normal or lognormal distribution-based likelihood function or the effective sample size of a multinomial distribution based likelihood function). Outliers can be addressed by using robust likelihood functions.

*W(Composition residuals)*

Methods that automatically estimate the data weighting should be used (the Francis method is currently recommended) and therefore the magnitude of the composition residuals does not have to be considered. Trends in residuals should be checked using plots of: 1) overall fits (although the "empirical" selectivity plots might be preferable, see below), 2) average age/size fits, 3) residuals vs time, 4) residuals vs age/length, 5) residuals by cohort (if age data), and 6) bubble plots of time and age/length. The weights should be based on any major issues found in any of these graphical methods to evaluate residuals.

*W(Index residuals)*

When the standard deviation of the likelihood function is fixed, it should be compared with the RMSE. Both the fit of the index and the raw residuals should be plotted to determine trends in the residuals.

*W(Recruitment residuals)*

Recruitment is typically modelled assuming a lognormal distributional assumption with a fixed standard deviation used in the associated penalty. The RMSE should be compared with the assumed standard deviation. Trends in recruitment are expected because recruitment is often driven by autocorrelated environmental processes. However, trends also may be related to a misspecified stock-recruitment relationship, and so residuals should be plotted against abundance (however care is needed in the interpretation because recruitment could drive abundance). Also, trends related to increases in catch may be an artifact of the model (see the bigeye tuna application below). The usefulness of a recruitment residual diagnostic is currently unclear.

#### 2.2.2.c Assigning and rescaling weights

Weights need to be assigned to each of the models (using the weight categories described above) following the hierarchy and based on evaluation of the metrics described above. The model weights must then be rescaled so that they represent probabilities and can be used in the risk analysis. There are two related factors that need to be considered but may not necessarily be treated the same: i) when should the weights be rescaled to sum to one, and ii) how to assign the weights for a specific model relative to the other models. Since the grouping of models and hypotheses prior to rescaling may dictate how models and hypotheses are grouped for assigning weights, we first discuss rescaling, even though when implementing these procedures clearly the weighting must be done before rescaling.

**i) when should the weights be rescaled to sum to one**

Rescaling conditionally, according to the branches of the hierarchy, which turns the flow chart into a probability tree, may be desirable. For example, without rescaling conditionally on the hierarchy, the more models that are used to represent a hypothesis, the more weight that hypothesis gets when generating

management advice. This could lead to overweighting of a specific hypothesis. To illustrate this point, we use the example of steepness of the Beverton-Holt stock-recruitment relationship fixed at 0.7, 0.8, and 0.9, in three different models. When all these models are given equal weight, the hypothesis that recruitment is related to spawning stock size, gets three times more weight than the alternative that recruitment is functionally independent of stock size (steepness fixed at 1.0). If additional models with steepness fixed at 0.5 and 0.6 were added, the weight would be 5 times higher, even if these additional values of steepness may be implausible for a particular stock. Of course, an approach for this example would be to estimate steepness and let the data determine the weight, but as discussed below (see text related to Level 3 above), estimating steepness is inappropriate in some cases, and particularly for steepness.

We propose the following:

1. Rescale the Level 1 overarching hypotheses weights to sum to one across all overarching hypotheses. These weights will then be multiplied by the weights from the other levels.
2. Rescale the Level 2 weights to sum to one within each sub-level (*e.g.* A, B, …) within a branch of the hierarchy (*i.e.* for a given Level 1 overarching hierarchy). The exception to this is for model fit when some models have less/different data or down-weighted data. In this case, subdivide the models further into groups of models with the same data and data weighting and rescale the weights for models in each of these groups to sum to one.
3. Rescale the Level 3 weights to sum to one within a branch of the hierarchy (*i.e.* for a given Level 2 hypothesis).

**ii) How to assign the weights for a specific model relative to other models**

As with rescaling the weights, in some cases it may be beneficial to follow the hierarchy when assigning weights, while in other cases it may be more appropriate to assign weights to a metric based on evaluation of all models or a subset of models on different branches. For those metrics that measure an aspect of model reliability that is branch-specific or are not based on data (*e.g. W(Expert)*), the hierarchical structure of hypotheses should be followed when assigning weights. For those metrics that measure an aspect of model reliability that has the same interpretation across all branches in the hierarchy (*e.g. W(Plausible results)*), the weights should be assigned globally.

We recommend the following:

1) Assign the Level 1 *W(Expert)* weights relative to all overarching hypotheses.
2) Assign Level 2 weights to *W(convergence)*, *W(Plausible parameters)*, *W(Plausible results)* and *W(Diagnostics)* relative to all models and hypotheses.
3) Assign Level 2 weights to *W(Fit)* relative to models that use the same data or data-weighting, which may correspond to several different branches in the hierarchy.
4) Assign Level 2 weights to *W(Expert)* relative to models in the same branch of the hierarchy (i.e. for a given Level 1 overarching hypothesis).
5) Assign Level 3 weights relative to models in the same branch of the hierarchy (*i.e.* for a given Level 2 hypothesis).

We suggest that the weighting be done using a panel of experts who discuss the various models with respect to each of the metrics. If no consensus on weights for a particular metric is obtained among the experts, then each expert can assign his/her own weight, and those weights are then averaged over experts to obtain a weight for that metric for a model. We recommend rescaling the weights to sum to one for each expert before averaging so that each expert's weights have more consistent influence.

Clearly, if a weight component is not relevant (e.g. the *W(Plausible parameters)* metric is not relevant if there is no parameter estimated that represents the hypothesis), it should be assigned to a value of 1.

#### 2.2.2.d 2d) Reducing the number of models and the analyses conducted

There are usually a variety of hypotheses and combinations of the hypotheses that can address the stock assessment issues and represent the state of nature. The number of models representing these hypotheses can easily get impractical to implement, particularly since some of the diagnostics (*e.g.* $R_0$ profile and retrospective analysis) need the model to be run several times. Therefore, an approach is needed to efficiently eliminate unlikely models. Per the formula for *W(Model)*, if any of the metrics is assigned a weight category of *None* (zero) then the model gets an overall weight of zero and is eliminated.

One of the benefits of creating the hierarchy of models and hypotheses is that it may allow elimination of groups of models without running all the models. This can be done by defining a "base" model from which the other models in the sub-levels of Level 2 are derived. The base model would typically be a simpler model (*e.g.* some parameters that are estimated in the other models are fixed) and if this model fails, then the other models derived from this model are also eliminated. Implementation of this approach needs to consider the reason for the elimination of the base model, because models derived from the base model may in fact correct for the reason the base model was eliminated.

It is also useful to check the categories *W(Expert) and W(Convergence)* early in the analysis because when these eliminate models, the calculations for the other categories do not need to be conducted, which will save time.

### 2.2.3. Calculating probability distributions for quantities of interest for a model

As mentioned above, a model with parameters representing alternative hypotheses run in a Bayesian framework would be preferable for estimating probabilities. However, at present, running MCMC on many contemporary assessments is impractical. The computational demands, potential for parameters estimated with low precision (*e.g.* selectivity parameters) and the correlation among parameters make for prohibitively long run times to get reliable posterior distributions. Models could be reparametrized to reduce parameter correlations, prior distributions could be added or parameters estimated with low precision could be fixed. However, this is not always possible, or doing so may be inappropriate, impractical, or time consuming for a large number of models. Therefore, it is necessary to use some approximations. First, we have to use frequentist approaches to approximate the posterior for the quantity based on a specific model *m*, $P(Quantity|Model{=}m)$. This can work well when the data is very informative since, in a Bayesian analysis, one characteristic of a good prior is that it has frequentist matching properties (*e.g.* the correct coverage probability), which typically improve with increasing sample size for common frequentist approaches. Second, the probability distribution may be asymmetrical, but conducting a profile likelihood for the desired quantity may not be possible (*e.g.* for a derived quantity). Therefore, normal approximations based on the estimated standard deviation (standard error) of a derived quantity are used. The resulting distribution is rescaled to obtain $P(Quantity|Model{=}m)$. Posteriors derived from limited MCMC analyses could be used to evaluate appropriateness of the approximation.

### 2.2.4. Combining probability distributions across models

The probability distribution function (PDF) for a quantity of interest, $P(Quantity)$, that corresponds to a collection of models needs to be obtained in order to evaluate the probability that the quantity of interest exceeds a reference point. We estimate $P(Quantity)$ using model-averaging. To compute $P(Quantity)$, $P(Quantity\ |Model{=}m)$ (from Section 3) is evaluated for each model *m* in the collection of models across an interval of regularly-spaced discrete values of the quantity of interest, and weighted by the rescaled

values of *W(model)*, which we will refer to as "*P*(*Model*=*m*)", before summing across models and rescaling. The spacing among values of the quantity of interest must be fine enough and cover a large enough range of values that the tails of *P*(*Quantity*) can be used to accurately assess tail probabilities [*e.g.* P($F > F_{LIMIT}$) > 0.1]. More formally, *P*(*Quantity*) is given by:

$$P(Quantity) = \sum_{m \in \{Models\}} P(Quantity|Model = m)P(Model = m)$$ **[Equation 6]**

The full algorithm used to compute *P*(*Quantity*) can be summarized as follows:

a) Determine the weight of each model in the collection (*i.e. W(model)* for each model, per Section 2).
b) Rescale the values from (a) to obtain *P*(*Model* = *m*) for every model in the collection.
c) Calculate the probability of the quantity of interest for each model across an interval of regularly-spaced discrete values (Section 3) and rescale so that they sum to one. This gives *P*(*Quantity* |*Model*=*m*).
d) Multiply (b) and (d) for each model in the collection and sum across models to give *P*(*Quantity*).
e) Evaluate (d) for all management quantities.

Finally, in order to evaluate the probability of exceeding a reference point, the cumulative distribution function (CDF) for the quantity is computed from *P(Quantity)* using the trapezoidal rule. First, estimate the probability for the interval between that value $q_i$ and the previous value, $q_{i-1}$,

by: $\int_{q_{i-1}}^{q_i} P(Quantity)\, dQ \approx \text{abs}(q_i - q_{i-1}) \frac{[(P(Quantity=q_i)+P(Quantity=q_{i-1})]}{2}$ **[Equation 7]**

The CDF is then obtained by summing up these values in series and rescaling by dividing by the maximum. When the normal approximation is used to represent the pdf of the quantity of interest, the cumulative normal distribution can be used to reduce computational demands.

### 2.2.5. Presenting the results in the form of a risk analysis

The results can be presented in a number of different formats depending on the purpose of the analysis. Plotting the PDFs of the quantities of interest (e.g. the ratio of the current status to the reference point) can be used to present the shape of the distributions. It is often useful to present the pdfs by components or a combination of components to illustrate the influence of the weighting factors. Cumulative density functions (CDFs) can be used to determine the probability of exceeding the reference points.

A variety of decision tables could also be created, but the most common is the outcome of specific management action under different states of nature. The states of nature could be the individual models, or if there are too many, combinations of models, or some derived quantity such as biomass level. The probability of each state of nature is usually also included in the table to help interpret the overall results integrated across all states of nature.

Sensitivity analysis to the weights is also useful to determine how robust the results are to uncertainty in the weight assignments.

## 3. BIGEYE TUNA APPLICATION

Bigeye tuna in the EPO is used to illustrate the risk analysis approach. In this section we discuss the hierarchy of hypotheses and models and the weighting system that were developed for this application.

*Hierarchy of hypotheses and models*

There are two main issues to address in the bigeye tuna application: the regime shift in recruitment and misfit to the composition data for the fishery that has asymptotic selectivity. To address these issues many

combinations of models were initially considered, and their graphical presentation gets a little disorderly due to the early elimination of some models based on the weight assignments (see below) and the desire to keep the number of models limited. Figure 2 shows the full set of models considered and Figure 3 shows those that were not eliminated. In the text, models and hypotheses are identified in *italic* font.

**Level 1**

The main unresolved issue in the bigeye tuna assessment is the estimated regime shift in recruitment that coincides with the expansion of the purse-seine fishery on floating objects. This could be a real regime shift in recruitment caused by changes in environmental conditions, predation, or competition, or it could be due to model misspecification. The recruitment regime shift becomes the basis for the overarching hypotheses: 1a) the *regime shift is real* and 1b) the *regime shift is an artifact* of model misspecification (Figure 2).

The issue of the recruitment regime shift forms the basis for the overarching hypotheses because available data cannot be used to clearly establish whether the recruitment regime shift is real. Obviously, adding a parameter representing a regime change in recruitment would easily explain the data, but this is a convenient fix and does not rule out model misspecification. Although the data used in the model cannot clearly differentiate between the two overarching hypotheses, they could be used to estimate parameters that represent the possible model misspecification.

**Level 2**

Within Level 2, hypotheses and corresponding models are grouped into sub-levels according to whether they address the regime shift in recruitment (Level 2A) or misfit to composition data for the fishery that has asymptotic selectivity (Level 2B). Some hypotheses may attempt to address both issues simultaneously, while in other cases multiple hypotheses may be needed (Figure 2).

**Level 2A**

In Level 2A are hypotheses that address the regime shift in recruitment. Two hypotheses are used to represent the overarching hypothesis that regime shift in recruitment is real, one representing a regime shift in the environment (*Environment*) and the other representing the purse seine fishery on floating objects reducing the level of predators or competitors (*Ricker*). Models under the Environment hypothesis estimate a parameter to represent the regime shift in recruitment, which allows the hypotheses on the next sub-level (Level 2B) to focus on processes that may improve the fit to the composition data without necessarily influencing the regime shift.

Nine hypotheses are used to represent the overarching hypothesis that regime shift in recruitment is an artifact of model misspecification. Five of these hypotheses are based on preliminary model runs that indicated the regime shift in recruitment can be reduced by increasing the biomass so that the catch does not have such a large impact on the abundance. In these preliminary model runs, the low proportion of large fish in length composition data for the longline fishery, which was assumed to have an asymptotic selectivity, kept the abundance low because more larger fish would have been expected with a higher biomass (lower fishing mortality). Decreasing the asymptotic average length (*Estimate growth*) or allowing the selectivity for this fishery to be dome-shape (*Dome selectivity*) allows the biomass to be larger. Increasing adult natural mortality (*Adult M*) or down-weighting the length composition data for the fishery with asymptotic selectivity (*Unreliable longline length composition*) also allows the biomass to be larger while predicting few large fish in the catch. These models also address the misfit to the length composition data. There is one more hypothesis in this group, (*Pre-adult movement*), which does not necessarily increase the biomass estimates but potentially addresses both the regime shift and composition misfit

simultaneously. This hypothesis is based on the assumption that there is movement of fish between the central Pacific Ocean (CPO) and the EPO that are of an age between that of the fish caught in the purse-seine fishery on floating objects and that of the fish caught in the longline fishery. Because the longline fishery CPUE is used to calculate the index of abundance, the abundance of these fish would not be represented by the index.

The other four hypotheses can reduce the recruitment regime shift without necessarily increasing the biomass, but do not address the composition misfit. These include the early catch being under-estimated or the later catch being over-estimated (*Misreported Catch*), the index of abundance not being representative of the stock (*Index not representative*), and spatial structure within the EPO (*EPO spatial structure*). The final hypothesis (*Short-term model*) addresses the regime shift in recruitment by assuming it is due to some unknown model misspecification in the early period (prior to 2000) that cannot be identified/resolved with available data, and thus, is not addressed by the other models.

**Level 2B**

On Level 2B are the hypotheses that address the fit to the length composition data for those hypotheses in Level 2A that are solely focused on addressing the recruitment regime shift issue. Specifically, on Level 2B are the hypotheses: *Estimate growth*, *Dome-shape selectivity*, and *Estimate adult natural mortality*. For the sake of completeness, a model without these changes is also included (labelled "*Fixed*"). As noted above, some of these hypotheses also potentially remove the regime sift in recruitment. This duality is indicated in Figures 2 and 3 by representing these hypotheses as vertical boxes that extend across both Level 2A and Level 2B.

**Level 3**

Finally, on the lowest level in the hierarchy (Level 3) are the models representing sub-hypotheses for different values of the steepness of the Beverton-Holt stock-recruitment relationship. Steepness is treated differently from the other model parameters because simulation work (*e.g.* Lee et al. 2012 ) has shown that steepness estimates within stock assessment models can be biased (particularly for higher productive stocks), and therefore it may be preferable to not use the data in the model to inform the value of steepness (with relatively unproductive stocks with good contrast in spawning biomass probably the exception). We run the models with different fixed levels of steepness so the fit to the data does not influence the weights assigned to models and hypotheses in Level 2. The weight for each value of steepness was only based on expert opinion and whether the model obtained a positive definite Hessian.

*Weighting system*

Weight metrics: new metrics

The main issues in the bigeye assessment prompted us to add two more weight metrics to the metrics shown in Section 2.2 that are used to calculate the final weight of each model (i.e., *W(model)*):

1) *W(Fix regime)*: The ability of the model to correct the regime shift in recruitment.
2) *W(Empirical selectivity)*: How well the assumed selectivity curves represent the implied selectivity.

In addition, several weight metrics needed modification to address specific issues in the bigeye tuna application and these are also discussed below.

W(Fix regime)

The *W(Fix regime)* weight represents the ability of the model to remove the regime shift in recruitment. The magnitude of the estimated regime shift is measured by the ratio of the median recruitment in the late regime (1994-2019) to the median recruitment in the early regime (1979-1993). The values of the regime shift metric, $R_{shift}$, that correspond to each weight category are given in Table 1. For the models related to the overarching hypothesis that the regime shift is real (*Environment*, *Ricker*), and for the *Short-term* hypothesis, the weight is set to 1.

W(Empirical selectivity)

The *W(Empirical selectivity)* weight represents the ability of the model to estimate the appropriate selectivities. Estimated and "empirical" selectivities should be similar (except for fisheries that are not fit to the composition data). The "empirical" selectivity is calculated by taking the average catch at length in numbers for a fishery and dividing it by the average numbers at length in the population, where the averages are evaluated across all years. Differences between the two indicate that the selectivity curve is inappropriate (*e.g.* too inflexible or assumed asymptotic but dome-shape is required to fit the data). In the case of asymptotic selectivity, misfit may indicate that either a dome-shape selectivity is needed, or when fitting to length composition data, that the growth curve is misspecified. The plot of estimated *versus* "empirical" selectivity provides a clearer visual picture of misfit than do plots of overall observed and predicted length compositions. This is because the latter gives less emphasis to large-sized fish, which are less abundant, whereas the former will clearly show misfit at large lengths, which can be strongly influenced by assumptions about selectivity.

*W(Empirical selectivity)* is treated as a separate weight metric in the bigeye application, rather than adding it to *W(Diagnostic)* because it addresses one of the major issues with the assessment, and because misfit to the composition data typically has a large impact on the stock assessment results. As with other diagnostic criteria, lower weights should be given to larger misfits.

*Weight metrics: modifications to original procedures*

The *W(Fit)* weighting procedure had to be modified because some of the models use different data sets or data-weighting. Specifically, the model that estimates growth (*Estimate growth*) includes the otolith data and therefore *W(Fit)* for this model was evaluated based on the difference in AIC obtained when the otolith data were excluded. The *Index not representative* model uses the catch-curve diagnostic from the *Environment* model and therefore does not fit to the index of relative biomass; however, this model was eliminated before it was necessary to evaluate *W(Fit)* so there was no need to modify the *W(Fit)* procedure for this model. The *Unreliable longline composition* model down-weights the survey and longline fishery length composition data; however, this model was also eliminated early in the model weighting process, prior to evaluation of *W(Fit)*. The *Short-Term* model has less data and therefore the AIC for these models were evaluated separately from the other models which all covered a longer time period (i.e., were medium-term models).

Maximum gradient information was not used to evaluate *W(Convergence)*. This is because it is not clear whether a large maximum gradient means that the model has not converged on the global MLE or whether the likelihood surface is flat and that it is appropriately reflected in the parameter uncertainty. Therefore, we assigned values to *W(Convergence)* based solely on whether the Hessian was positive definite.

*W(Diagnostics)* was only evaluated for the sub-models in Level 3 with steepness of $h$=1.0. The steepness runs ($h$ = 0.7, 0.8, 0.9, 1.0), which were applied to all models in Level 2, except the *Ricker* model, greatly increased the number of model runs. It was not feasible to run all the diagnostics for all of these models, so for each of these models we assumed that the diagnostics for sub-models with steepness values other

than 1.0 would be similar to those of the sub-model with steepness *h*=1.0. If sub-model with a steepness of *h*=1.0 was eliminated due to any of the weighting factors (assigned a value of *None*), then the sub-models with different values of steepness were not run. The only weighting metrics used for the sub-models with steepness less than 1.0 were *W(Expert)* and *W(Convergence)*.

Only *W(Expert)* was used to weight the overarching hypotheses of Level 1. Weights for all the models on the lower levels (Levels 2-3) were assigned *W(Expert)* conditional on the overarching hypothesis to which they belong.

*Assigning weights*

Weights were assigned by six experts. The results are presented for each expert (see below) but are randomized for each metric to obscure the expert. Due to the subjective nature of many of the weighting metrics, there was not always agreement on weights among the experts. There was consensus on some weighting assignments, but in most cases there were differences in the weighting assignments. Therefore, we took the approach where the weighting assignments were first discussed among all experts and then each expert provided their own weighting for each of the metrics. The average of the individual weighting assignments, first scaled to sum to one for each expert, were used as the final weighting assignments to compute the model weights.

Weights were assigned according to the hierarchy of hypotheses shown in Figure 2. The sub-levels in Level 2 of the hierarchy (i.e., Levels 2A and 2B) allow the assignment of weights for some metrics (*e.g. W(Expert)*) to models on Level 2a independent of Level 2b, and the model weight is the product of weights given to models on the two sub-levels. However, this is not possible for the models where the same hypotheses are used to solve both the recruitment regime shift and the composition data misfit, but they were still assigned weights independently for Level 2a and Level 2b.

The method to reduce the number of models tested was based on selecting a "base" model for each Level 2A hypothesis and if it was eliminated by receiving a weight of *None* for any metric then the other models based on that hypothesis were not conducted. The "base" model fixed parameters that were related to Level 2B and not Level 2A (i.e. the models that used the same hypothesis for Level 2A and 2B had the respective parameters estimated) and steepness equal to one. Several hypotheses were eliminated early in the assignment of weights. The *Ricker* model with fixed growth, fixed natural mortality, and asymptotic selectivity (*Fixed*) did not have a positive definite Hessian and no further Ricker models were run. The *Index not representative* model did not improve the recruitment regime shift. The *EPO spatial structure* and *Misreported catch* models were assigned the *None* weight category for *W(Expert)*. The *Unreliable longline composition* model was assigned the *None* weight category for *W(Empirical selectivity)*. The modified hierarchy flowchart without these models is shown in Figure 3. However, it is not certain that these hypotheses that were eliminated early in the weighting process would also have been eliminated if combined with other hypotheses at Level 2B.

*Calculation of reference points*

The IATTC HCR for tropical tunas in the EPO that applies to bigeye tuna is based on both fishing mortality and spawning biomass targets and limits. The targets are the fishing mortality ($F$) and spawning biomass (calculated dynamically; $S_d$) that correspond to maximum sustainable yield (MSY). The MSY quantities are based on the average fishing mortality over the last three years in the assessment. $S_{MSY\_d}$ can be conceptualized as projecting the model forward under the historical estimated recruitment using $F_{MSY}$. The limit reference points correspond to the spawning biomass where recruitment is reduced to 50% based on a Beverton-Holt stock-recruitment relationship with a steepness ($h$) of 0.75 (Maunder and Deriso

2014). The spawning biomass limit reference point ($0.077S_0$) is based on equilibrium spawning biomass (using average recruitment, adjusted for the stock-recruitment relationship, over the modelling period) not the dynamic spawning biomass. The fishing mortality limit reference point is the fishing mortality corresponding to this spawning biomass. The evaluation of the current status relative to the reference points is done using the following ratios: $S_{cur}/S_{MSY\_d}$, $S_{cur}/0.077S_0$, $F_{cur}/F_{MSY}$, and $F_{cur}/F_{0.077S0}$, where $S_{cur}$ is the spawning biomass at the start of 2020 and $F_{cur}$ is the average fishing mortality from 2017 to 2019.

Stock Synthesis (SS version 3.30.15 and modifications provided by Richard Methot (NOAA Fisheries) to calculate the standard deviations for the management quantities) are used to conduct the stock assessment for bigeye tuna (see *Model descriptions* section below for more details). However, obtaining the standard deviations for the ratios of the current status to the reference points is not always straightforward in SS. Thus, in the following we describe the methods used to calculate the standard deviations.

The four quantities of interest are:

**$S_{target}$: $S_{cur}/S_{MSY\_d}$**

The expected value of $S_{cur}/S_{MSY\_d}$ is calculated by projecting the model forward under $F_{MSY}$ using the recruitment deviates from the historic period (appropriately adjusted using the bias correction ramp). The standard deviation is not available for this quantity and therefore the CV is assumed to be the same as for $\boldsymbol{F_{cur}/F_{MSY}}$.

**$S_{LIMIT}$: $S_{cur}/0.077S_0$**

The standard deviation of $S_{cur}/S_0$ is available from Stock Synthesis and based on $Var[cX] = c^2\ Var[X]$, $Var[S_{cur}/0.077S_0] = (1/0.077)^2 \times Var[S_{cur}/S_0]$

**$F_{target}$: $F_{cur}/F_{MSY}$**

$F_{cur}/F_{MSY}$ and its associated SD is available from Stock Synthesis.

**$F_{LIMIT}$: $F_{cur}/F_{0.077S0}$**

$F_{cur}/F_{0.077S0}$ and its associated SD is available from Stock Synthesis (using the target biomass denominator option in the starter file).

*Decision table*

We provide a decision table presenting the outcome of different levels of fishery closures. We assume that the fishing mortality is proportional to the length of time the fishery is open, adjusted for changes in fishing capacity. This assumption can be used to determine the fishing mortality with a different closure compared to the current estimated fishing mortality under the current closure. The effective length of time the fishery is open is adjusted for the effect of the Corralito.

Decision tables usually present the outcomes of alternative management actions. In the bigeye tuna example, the management action is the number of days of closure and the outcome is the probability of the fishing mortality rate being below that corresponding to MSY ($F_{MSY}$) or the limit ($F_{LIMIT}$). The calculation is not straightforward because: a) the closure must be converted into a fishing mortality rate, and b) the only information available to calculate the quantity of interest is a probability distribution for $F_{cur}/F_{MSY}$ (or $F_{cur}/F_{LIMIT}$). The calculations are therefore made based on the ratio of the F associated with the desired closure ($F_{new}$) to $F_{cur}$. We assume that the fishing mortality is proportional to the amount of time the fishery is open (365 - closure) adjusted by the change in fleet capacity and the corralito: $F = q \times (365 - (\text{Closure} + \text{Corralito})) \times \text{Capacity}$, where q is an unknown scaling constant. Calculating the ratio $F_{new}/F_{cur}$ eliminates the scaling factor:

$$\frac{F_{new}}{F_{cur}} = \frac{(365 - (\text{Closure}_{new} + \text{Corralito})) \text{ x Capacity}_{new}}{(365 - (\text{Closure}_{cur} + \text{Corralito})) \text{ x Capacity}_{cur}} \equiv c \textbf{ [Equation 8]}$$

where $\text{Capacity}_{\text{new}}$ is the well capacity in cubic meters for 2020 and $\text{Capacity}_{\text{cur}}$ is the average well capacity over 2017 to 2019, Corralito is the equivalent number of days closure represented by the Corralito spatial closure (3 days), and $\text{ClosureDays}_{\text{cur}}$ is the average number of days of closure in 2017 to 2019 (72).

Rewriting the above equation and dividing both sides by $F_{MSY}$ we have:

$$\frac{F_{new}}{F_{MSY}} = c\frac{F_{cur}}{F_{MSY}} \equiv g\left(\frac{F_{cur}}{F_{MSY}}\right) \textbf{ [Equation 9]}$$

If we assume that *c* is a constant, we can represent the unknown quantity, $\frac{F_{new}}{F_{MSY}}$ , in terms of $\frac{F_{cur}}{F_{MSY}}$ multiplied by the constant *c*, which for simplicity we will label as the function "*g*". Because the function *g* is strictly increasing its inverse is well-defined and we can use the method of transformations for CDFs to obtain the probability that $\frac{F_{new}}{F_{MSY}}$ is less than or equal to 1:

$$P\left(\frac{F_{new}}{F_{MSY}} \leq 1\right) = P\left(\frac{F_{cur}}{F_{MSY}}c \leq 1\right) = P\left(\frac{F_{cur}}{F_{MSY}} \leq \frac{1}{c}\right) \textbf{ [Equation 10]}$$

*Model descriptions*

Stock Synthesis (Methot and Wetzel 2013) version 3.30.15 is used to implement catch-at-length age-structured integrated models for the latest (2020) stock assessment of bigeye tuna in the EPO (Xu et al. 2020). Fisheries are defined by area (geographic subdivisions of the EPO), fishing method (longline, purse-seine), and purse-seine set type (on floating objects, on tunas associated with dolphins, and on unassociated schools of tuna). Typically, models are fitted to indices of abundance based on longline CPUE data and to length-composition data associated with the indices of abundance and the fisheries. When growth is estimated, the model is also fitted to conditional age-at-length composition data derived from otoliths. See Xu *et al*. (2020) for a full description of the models and results.

A description of each hypothesis at Level 2A in the hierarchical flow chart (Figure X), including those that were eliminated during the assignment of weights (compare Figure Y to Figure X), and a brief rationale for the various hypotheses, are provided below.

*Base reference model*: This model is the basis for all other models and is not used in the risk analysis. The selectivity for one of the longline fisheries is asymptotic. This model is similar to the base case model used in previous assessments except that the weighting for the composition data uses the Francis method.

*Environment*: This model assumes that the regime shift is real and is caused by a change in the environment. The selectivity for one of the longline fisheries is asymptotic. This model is similar to the base case model used in previous assessments except that the weighting for the composition data uses the Francis method and it estimates a parameter representing the change in recruitment.

*Ricker*: This model uses a Ricker stock-recruitment relationship and assumes that a reduction in predators caused the increase in recruitment. The model assumes that mature bigeye tuna are the predators (cannibalism), but the bigeye mature biomass can also be assumed to be a proxy for other tunas and predators impacted by the expansion of the purse-seine fishery on floating objects (*e.g.* skipjack, yellowfin, sharks) that may consume juvenile bigeye. This model would not converge resulting in a Hessian that was not positive definite and a large maximum gradient. Therefore, although it is included in the flow chart, it was assigned a zero weight for *W(Convergence)* and not included further in the calculations.

*Index not representative*: The index from the longline fishery CPUE data does not show a substantial decline when the floating-object fishery expanded and therefore may not be proportional to abundance

or may represent the abundance of bigeye from a stock other than the one that is caught in the floating-object fishery. This model is based on the catch curve diagnostic of the *Base reference* model, which does not use the index of abundance.

*Spatial structure within the EPO*: The longline index of abundance may be representing an area different from that into which the floating-object fishery expanded. Many models were run investigating this hypothesis (Valero *et al*., 2018; Valero *et al*. 2019a; Valero *et al*., 2019b) and none could explain the regime shift in recruitment. Therefore, although it is included in the flow chart, it was assigned a zero weight for *W(Expert)* and not included further in the calculations.

*Misreported Catch*: Higher catch in the historic period would require higher recruitment and reduce the recruitment regime shift. Lower catch in the floating-object fishery in the later period would not require as much of an increase in recruitment, reducing the regime shift. Models that increased historical catch or reduced recent catch had to make unrealistically large change in catch to explain the recruitment regime shift. Therefore, although it is included in the flow chart, it was assigned a zero weight for *W(Expert)* and not included further in the calculations.

*Short-term model*: This hypothesis is evaluated using only the data from 2000 – 2019. It is assumed that the regime shift in recruitment is due to some unknown model misspecification in the early period (prior to 2000) that cannot be identified/resolved with available data, and thus, is not addressed by the other models.

*Pre-adult M (movement)*: This model approximates movement of fish into or out of the EPO from the central Pacific Ocean (CPO) by applying natural mortality to fish starting at an age that is between those selected by the floating-object fishery and those selected by the longline fishery. Higher natural mortality represents fish moving out of the EPO and lower natural mortality represents fish moving into the EPO. This modified mortality schedule also could capture actual differences in age-specific natural mortality driven by a variety of processes.

*Estimate growth*: Estimating growth allows for a larger biomass and therefore reduces the regime shift in recruitment. The length composition data for the fishery with asymptotic selectivity has few fish around the asymptotic length and therefore the model estimates a high fishing mortality, and corresponding low biomass, to reduce the number of large fish and fit the length composition data. Estimating growth produces a low asymptotic length, reducing the predicted number of large fish and fits the length composition data without increasing fishing mortality, which allows for a larger biomass. All four parameters of the Richards growth curve and the two parameters representing the variation of length at age are estimated. The model is fit to the otolith age conditioned on length data. This model can also address the misfit to the length composition data at Level 2B.

*Dome selectivity*: A dome-shape selectivity for the longline fishery allows for a larger biomass and therefore reduces the regime shift in recruitment. The length composition data for the fishery with asymptotic selectivity has few fish around the asymptotic length and therefore the model estimates a high fishing mortality, and corresponding low biomass, to reduce the number of large fish and fit the length composition data. Estimating a dome-shape selectivity reduces the predicted number of large fish caught allowing the model to fit the observed length composition data, but also produces a “cryptic biomass” increasing the biomass estimate. A double normal selectivity curve is used. This model can also address the misfit to the length composition data at Level 2B.

*Adult M*: Estimating adult natural mortality allows for a larger biomass and therefore reduces the regime shift in recruitment. An increased value of natural mortality reduces the fishing mortality that is needed

to fit the length composition data and therefore increases the biomass for a given level of catch. This model can also address the misfit to the length composition data at Level 2B.

*Unreliable longline composition*: Down-weighting the longline composition data allows for a larger biomass and therefore reduces the regime shift in recruitment. Down weighting the longline composition data frees up the model from having to keep the biomass low to fit the lack of observed large fish and therefore estimates a larger biomass to reduce the recruitment regime shift. Although, it is included in the flow chart, it was assigned a zero weight for *W(Empirical selectivity)* and not included further in the calculations.

## 4. RESULTS

The following provides the results of the Risk analysis and are based on the results from the bigeye tuna stock assessments that can be found in Xu et al. (2020). We present the weight categories assigned and the consequent probabilities that are scaled to sum to one within an overarching hypothesis. When presenting these results, we use the abbreviations shown in Table 2 to refer to the various models of Level 2 (Figure 2).

### 4.1. Assigning weights

#### 4.1.1. Level 1: Is the regime shift real?

There was consensus among the experts that the probability of the recruitment regime shift being real was low. Either a change in the ecosystem occurring at the same time as the increase in catch of the floating-object fishery has to be a coincidence or the increased catch changed the ecosystem. There have been some physical and biological changes in the pelagic EPO (see IATTC 2013), but their timing and magnitude does not necessarily correspond to the increase in bigeye recruitment and similar recruitment patterns are not observed for yellowfin tuna in the EPO. There is some evidence of tuna eating juvenile tunas, but the consumption rates are low. Assessments of tropical tunas in other oceans have also found increases in estimated recruitment that correspond to expanding floating-object fisheries. Therefore, based on the fact that the recruitment increased when the floating-object fishery expanded and the lack of evidence of a corresponding ecosystem effect, the regime shift being real was given a *Low* weight and that it is an artifact of the modelling was given a *High* weight.

#### 4.1.2. Level 2

*W(Expert)*

The weight based on expert opinion is assigned before the results are considered for that model. There was some confusion of whether the weights should be assigned with consideration of the other models on the same level and branch of the flow chart or considering all models. It was concluded that they should be considered within a level and branch of the hierarchy (i.e. separately for each overarching hypothesis and issue being addresses). Describing the rationale for the weights by each expert and each model would be too voluminous to include here, so the general support for a model is included above in their description. *W(Expert)* was divided into Level 2A and Level 2B and was assigned based on whether the model addresses the regime shift in recruitment or the misfit to the composition data, respectively. The weights for the two components were multiplied and then rescaled within the overarching hypotheses for each expert before averaging. The models that were not run for combined hypotheses at Level 2B (*Ricker*, *Index*, and *Composition*) were given a weight of High for Level 2B. These models were eliminated from the analysis based on other weighting metrics and therefore this decision would have no impact on the results (they do influence the scaling of the weights to sum to one, but the relative size of the weights would remain the same).

Experts generally agreed on the *W(Expert)* weighting for each model, but some weight assignments differed by two categories (Table 3).

*W(Convergence)*

Weights were only assigned based on the Hessian being positive definite. All models (with $h$ = 1) had a positive definite Hessian and were given a *High* weight except for the *Ricker* model (Table 4). For some models at other values of $h$, the Hessian was not positive definite: the $h$ = 0.7, 0.8, and 0.9 steepness runs for the *Environment* model and the *Index unrepresentative* model, and $h$ = 0.7 for the *Short-Fixed* model (Table 4). The maximum gradient components were low for most of the runs where the Hessian was positive definite.

*W(Fit)*

The *Short-Growth* and *Environment-Growth* models fit their respective data best based on AIC, and *Short-Fixed* and *Movement* models fit their respective data the worst (Table 5). The *Index unrepresentative* model, which does not fit to the index of abundance, and the *Longline composition unrepresentative* model, were not included because they have different data or weighting and were eliminated due to other weighting metrics. *W(Fit)* was scaled to sum to one within models fitting to the same data/same data-weighting and within each of the overarching hypotheses.

*W(Plausible parameter estimates)*

There were differences among experts about the plausibility of the estimates of parameters representing the hypotheses, with some weight assignments differing by two categories (Table 6). The CVs on the parameters were low and therefore the uncertainty of the estimates was not taken into consideration. The pre-adult natural mortality was estimated to be about 0.1 (CV = 0.06) higher indicating an additional natural mortality of about 10% or 10% movement out of the EPO per quarter. This was considered reasonable by all experts, giving it a *High* or *Medium* weight. The estimate of the average length of the oldest age represented in the model was reduced from 196 cm to below 170 cm for the *Growth* and *Environment-Growth* models. This is somewhat inconsistent with the tagging data, which has the oldest 7 recovered fish having a length greater than 17 cm, but is consistent with historic length composition data. Given that the tagging data is limited in space and time, all experts considered the estimate reasonable and gave it a *Medium* or *High* weight. The *Short-Growth* model estimated average length of the oldest age at 184 cm, which is more consistent with the tagging data, and this model was also given *High* and *Medium* weights. For the Selectivity and *Environment-Selectivity* models, the estimated dome shape of the longline selectivity, which was considered asymptotic in other models, was very low for the oldest age represented in the assessment . Despite there being evidence that older fish spend some time deeper than the longline gear, this extent of doming of the selectivity curve was considered unreasonable by some of the experts, while others though it was reasonable, resulting in a mix of *Medium* and *Low* weights (Table 6). The estimated doming in the *Short-Selectivity* model was less and therefore was given *High* and *Medium* weights. The estimated adult natural mortality was double or more that estimated for the *Mortality* and *Environment-Mortality* models. It did not change much for the *Short-Mortality* model. The experts could not agree on the plausibility of these parameters, with weights ranging from *High* to *Low*. All the *Environmental* models estimated an increase in the recruitment and therefore this did not influence the weighting. The models that do not have parameters representing the Level 2A or 2B hypotheses (the *Fixed* models, *Index unrepresentative*, and *Longline composition unrepresentative*) are assigned a *High* weight.

*W(Plausible results)*

The *W(Plausible results)* metric was based on the fishing mortality and initial conditions (estimate of equilibrium catch and the parameter that scales the equilibrium recruitment). One way to remove the recruitment regime shift is to estimate a high biomass so that the expansion on the floating-object catch does not have an impact on the abundance. However, this may cause the biomass to be too high. The plausibility of biomass levels is hard to judge, and fishing mortality levels, which are related to the biomass level, may be easier to judge. The initial depletion level is also related to the estimated biomass level, so we compare the predicted initial equilibrium catch (and other associated parameters, initial recruitment offset and recruitment deviates) with historical catch to make sure they are somewhat consistent. The model is not fit to the equilibrium catch and there is only limited composition data for old fish in the early years, so the models based on a medium time frame cannot differentiate between initial fishing mortality and reduced initial recruitment. The *Short-term* model starts in a different year and therefore the initial catch levels will be different than the other models, and there is substantial length composition data at the start of the model to differentiate between initial catch and initial recruitment.

The experts generally agreed on most weights for the fishing mortality, although a few weight assignments differed by two categories, and they agreed on all the weights for the initial conditions (Table 7). The estimates of fishing mortality for the *Index unrepresentative* model were assumed unrealistically high and this model was given a weight of *None*. The *Growth*, *Selectivity*, and *Environment-Selectivity* models had extremely low fishing mortality for older fish and were given a weight of *Low* or *Medium*. The other models were generally assigned weights of *Medium* and *High*.

It was difficult to judge the initial depletion tradeoff between initial catch and initial recruitment. All models estimated a reduced initial recruitment except for the *Growth* model and three of the *Short* models. Only the *Short* models and the *Growth*, *Environment*, and *Environment-Mortality* models estimated equilibrium catch for the longline or purse seine initial "fisheries". The *Index* model was given a *Low* weight because the initial recruitment was substantially reduced, in addition to estimating initial catch, and this was thought to result in an unrealistically highly-depleted population. The *Growth* model was given a *Low* weight because the recruitment was not reduced and no initial catch was estimated, which resulted in an unrealistically undepleted population. The *Short* models were given a *Medium* weight because their estimated initial catch was substantially higher than the recent historical catch. The *Environment-Fixed* model was given a *Medium* weight because the initial recruitment is substantially reduced, and initial catch is estimated for both the longline and purse seine initial "fisheries".

*W(Diagnostics)*

Many of the diagnostics were similar across the alternative models and therefore it was decided that only the $R_0$ likelihood component profile, ADPM, and retrospective diagnostics would be used. The $R_0$ likelihood component and ADPM are combined based on the algorithm presented in Figure 1. When using the algorithm, a lack of information in the abundance index about $R_0$ was considered the same as no conflict between the abundance index and composition data. Several of the ASPM-Dev models did not converge and it was concluded that this meant that there was insufficient information, which was assumed equivalent to wide confidence intervals. All experts agreed on the weighting assignments. The *Index unrepresentative* model did not fit to the index of abundance and was given a default weight of *High*. This decision was not influential because the model was eliminated based on another weight metrics.

The estimated recruitment was similar for all models and the main feature was the regime shift that is already taken into consideration in the *W(Fix regime)* metric. The recruitment residuals, which are around the stock-recruitment relationship, will change for different values of steepness, but weights based on these are assumed to be included in the *W(h)* component. The composition residuals show trends over time and with length, but are similar for all models and are consistent with residual patterns seen in most

assessments. The main concern are the misfits to the large fish (175-200 cm) in the length composition data for the fishery with asymptotic selectivity, but that is covered in the *W(Empirical selectivity)* metric. The index residuals are similar among all models except for the *Longline composition unrepresentative* model, but this made the "empirical" selectivity even more dome-shaped and was given a *W(Empirical selectivity)* of *None* (see below), so the *W(Diagnostics)* is not important for this model. The catch curve diagnostic is very similar for all models, but does show a possibly change in selectivity in recent years that has not been addressed. The low biomass in the early period estimated by the Catch Curve Diagnostic is due to limited or no composition data in that period.

In all the models, except *Environment-Fixed* and *Selectivity*, the $R_0$ likelihood component profile showed that the composition data was driving the absolute abundance estimates. In all these cases, except the *Movement* model, the composition data was consistent with the index data, often because the index data had little information on absolute abundance. Since bigeye has variable recruitment, the ASPM-Rdev diagnostic was used. The *Environment* and *Short* ASPM-Dev models did not converge and they were assumed to have wide confidence intervals and therefore given a *Medium* weight. The other models had narrow confidence intervals. Of these, the *Movement* and *Selectivity* models had inconsistent biomass levels between the full model and the ASPM-Rdev model and were given a *Low* weight. The other models were given a *High* weight.

The retrospective analysis of spawning biomass ratio showed little retrospective pattern or variation for most of the models. Three *Environment* models were given *Medium* weight because the absolute scale varied with the retrospective runs. The *Movement* model was given *Medium* weight because the final estimate varied. The remaining models were given a *High* weight. A summary of the weight assignments for the components of *W(Diagnostics)* are shown in Tables 8 and 9 below.

*W(Fix regime)*

The probabilities were only based on the models with steepness equal to one. All the models, other than those that assumed the recruitment regime shift was real (*Environment*, *Ricker*) or did not model it (*Short-term*), which were given a default weight of *High*, reduced the regime shift to some degree except for the *Index* model, which was given a weight category of *None* (Table 10). The other models were given a weight of *Medium*, based on Table 1, except for the *Growth* model which was assigned a weight of *High*. The implied weight would be increased from *Medium* to *High* for some lower values of steepness, although these were not used in the risk analysis.

*W(Empirical selectivity)*

The experts all agreed on the weighting based on *W(Empirical selectivity)*. The consistency between the assumed selectivity curve and the "empirical" selectivity was generally good, except for the longline fishery that was assumed to have asymptotic selectivity. This fishery was the focus of this weighting metric. All models improved the consistency for this fishery over the *Environment* model except the *Longline composition unrepresentative* model. Because this fishery had down weighting of the composition data it was expected to fit worse, but because the dome shape of the "empirical" selectivity was even more extreme than the *Environmental-Fixed* model, this model was given a weight category of *None*. The models that estimated a dome shaped selectivity for this fishery (*Selectivity*, *Environment-Selectivity*, and *Short-Selectivity*) and the *Short-Mortality* model provided the best consistency and were given *High* weight, the remaining models were given *Medium* weight (Table 11).

#### 4.1.3. Level 3: Weights for steepness

As noted above, we use Level 3 in the hierarchy of models and hypotheses (Figure Y) to address steepness. The experts made their weight assignments for each level of the steepness parameter. The weights were

assumed the same independent of model. Weight assignments were then standardized within each expert and then averaged across experts.

There was large variability in the experts' opinions about weights for different values of the steepness of the Beverton-Holt stock-recruitment relationship. This parallels the general disagreement in the value of this parameter among fisheries scientists. Some experts gave *High* weight to high levels of steepness, which is consistent with the approach used in previous assessments of tropical tunas in the EPO, while others gave *High* weights to lower values of steepness, which is more consistent with other tuna RFMOs. One expert did not provide weights. The average of the weights across experts gave an almost linear positive relationship between the weight and the steepness value (Table 12).

### 4.2. Risk assessment results

The risk analysis for EPO bigeye tuna is used to evaluate several management quantities related to the HCR. These are grouped into 1) Current status as a ratio of the reference points, 2) Probability of exceed the reference points, and 3) Alternative management measures**.** These are all based on the probability distributions of the quantity of interest for each model and the model weights, which are presented first.

#### 4.2.1. Model probabilities

The probabilities are first assigned conditional on the overarching hypotheses about the recruitment regime shift. The model probabilities conditional on the overarching hypotheses are then multiplied by the probability of the overarching hypotheses to give the final model probability in Table 1, which are integrated over steepness. The steepness probabilities are the same for all models. However, several models did not have a positive definite hessian and these models were not included in the analysis and the probabilities were rescaled to sum to one across models within an overarching hypothesis. Therefore, the hypotheses that these models received a lower model probability. There is a range in probabilities assigned to the models with Short-Growth and Growth getting over 20% probability each. Several of the models got very low probabilities (less than 2 %).

#### 4.2.2. Probability distributions

$F_{cur}/F_{MSY}$

The combined distribution of $F_{cur}/F_{MSY}$ is bimodal (Figure 4). The bimodality is due to the substantial differences in the estimates from the **Short** models, which are more pessimistic ($F_{cur}/F_{MSY}$ mostly above 1), and the medium term models that do not assume the regime shift in recruitment is real (**Growth**, **Selectivity**, **Mortality**, and **Movement**), which are more optimistic ($F_{cur}/F_{MSY}$ mostly below 1). The remaining models (**Environment**), which assume the recruitment regime shift is real, fall in between these two sets of models (still with most of the mass in the optimistic side, $F_{cur}/F_{MSY}$ mostly below 1), but were assigned less weight. There is a substantial amount of the combined distribution above one indicating that the probability of $F_{cur}$ being above the target fishing mortality reference point is not negligible (50%, Table 13).

The hypotheses to explain the misfit to the longline composition data (**Growth**, **Selectivity**, **Mortality**) also have a large impact on the probability distribution with **Growth** and **Selectivity** being more optimistic and also having been assigned highest weights (Figure 4).

Steepness of the Beverton-Holt stock-recruitment relationship also influences the distributions of $F_{cur}/F_{MSY}$ with higher steepness being more optimistic, as expected.

***$F_{cur}/F_{LIMIT}$***

The combined distribution of ***$F_{cur}/F_{LIMIT}$*** is also bimodal and similar to the distribution for ***$F_{cur}/F_{MSY}$*** with the distribution, but shifted to the left (Figure 5). The composition of the model distributions is similar to the distributions for ***$F_{cur}/F_{MSY}$***, as expected. There is little probability (5%) above one indicating that the probability of ***$F_{cur}$*** being above the fishing mortality limit reference point is low.

***$S_{cur}/S_{MSY}\_d$***

The probability distribution for ***$S_{MSY}\_d$*** is generally bimodal, but also has some smaller modes (Figure 6). The small modes are because the individual model distributions for *S* are more separated than for *F*, but the CV used is based on the CV for ***$F_{cur}/F_{MSY}$***. The composition of the distribution is similar to that for *F*, but reversed on the X-axis. There is a substantial amount of the distribution below one indicating that the probability of the spawning biomass being below the target biomass reference point is high (at least 53%).

***$S_{cur}$ /$S_{LIMIT}$***

The probability distribution for ***$S_{cur}$ /$S_{LIMIT}$*** is bimodal and similar to the distribution for ***$S_{cur}/S_{MSY}\_d$***, but without the smaller modes and shifted to the right (Figure 7). The composition of the distribution is similar to that for ***$S_{cur}/S_{MSY}\_d$***. There is little of the overall model distribution below one indicating that the probability of exceeding the spawning biomass limit reference point is low (below 6%).

#### 4.2.3. Current status relative to reference points

Current status relative to reference points was calculated in two ways.

1. Take the point estimate of the ratio from the alternative stock assessment models and then use model averaging weighted by the model weights to calculate the ratio.

$$x = \sum_m P(model)\hat{x}_m \quad \textbf{[Equation 11]}$$

Where $\hat{x}_m$ is the maximum likelihood estimate (MLE) of the quantity of interest for model *m*

2. Use the expected value by integrating over the quantity of interest and weighting by the model weights

$$\bar{x} = \sum_m P(model) \int x\, P(x|model)dx \quad \textbf{[Equation 12]}$$

However, since the normal distribution, which is used as an approximation, is symmetric, these two methods give the same answer. If the distributions were not symmetrical, as would be expected with a better approximation, then the answers would differ.

Table 13 shows the ratio of the current status to the reference points for each model, averaged over steepness, and the value for all models combined. The results follow the central tendency of the distributions, as expected. $F_{cur}$ is 7% above $F_{MSY}$, $S_{cur}$ is 9% above $S_{MSY}$, $F_{cur}$ is well below $F_{LIMIT}$, and $S_{cur}$ is well above $S_{LIMIT}$.

#### 4.2.4. Probability of exceeding reference points

The probability of exceeding the reference points are calculated using cumulative distribution functions (CDFs). P($F_{cur}$>$F_{MSY}$) generally shows two groups of models corresponding to the modes in the probability distribution and their composition as noted above (Table 13). One group with a high probability of being below $F_{MSY}$ and one group having a low probability of being below $F_{MSY}$. The combined distribution has a 50% probability of $F_{cur}$ being above $F_{MSY}$.

The results for P($F_{cur}$>$F_{LIMIT}$) are like those for P($F_{cur}$<$F_{MSY}$), except in a few cases. The combined distribution has only a 5% probability of being above $F_{LIMIT}$.

The results for P($S_{cur}$ <$S_{MSY\_d}$) are like those for P($F_{cur}$>$F_{MSY}$), as expected. The combined distribution has a 53% probability of being below $S_{MSY}$.

The results for P($S_{cur}$ <$S_{LIMIT}$) are similar to P($F_{cur}$/$F_{MSY}$ ), as expected. The combined distribution has only a 6% probability of being below $S_{LIMIT}$.

#### 4.2.5. Alternative management measures

The risk analysis was used to determine the probability of exceeding the fishing mortality reference points for different days of closure. No projections were conducted so the spawning biomass reference points could not be evaluated. The fishing mortality is assumed to be proportional to the number of days the fishery is open adjusted by the Corralito spatial closure and changes in capacity. The fishing mortality reference points are evaluated under 6 different days of closure

1) Zero days of closure (0)
2) Half the current days of closure (36)
3) The days of closure that would give P($F_{cur}$ > $F_{MSY}$) = 0.5 (70)
4) The current days of closure (72)
5) The days of closure needed to achieve FMSY based on the expected value of $F_{cur}$/$F_{MSY}$ (88)
6) 100 days of closure

P($F_{cur}$ > $F_{MSY}$) is provided in Table 14. The sensitivity to the number of days of closure differs among models. This sensitivity depends on how close the probability distribution under the current days of closure is to 1. For example, three of the **Environment** models are very sensitive. Obviously, increased days of closure have a lower P($F$>$F_{MSY}$). The days of closure that would give P($F_{cur}$ > $F_{MSY}$) = 0.5 is very similar to the current days of closure. However, the days of closure needed to achieve $F_{MSY}$ based on the expected value of $F_{cur}$/$F_{MSY}$ is 16 days longer.

P($F_{cur}$>$F_{LIMIT}$) is provided in Table 15. The sensitivity to the number of days of closure also differs among models. This sensitivity depends on how close the probability distribution under the current days of closure is to 1. For example, all the **Short** models are very sensitive. Obviously, increased days of closure have a lower P($F_{cur}$>$F_{LIMIT}$).

## 5. DISCUSSION

### 5.1. Introduction

We have developed an approach to implement reference point-based fishery harvest control rules within a probabilistic framework that considers multiple hypotheses. The main features of this approach are: 1) hypotheses about states of nature are represented by alternative stock assessment models with specific model structure, data use and parameters; 2) hypotheses are grouped into a hierarchical framework, which highlights similarities among models thereby avoiding that any one hypothesis, or overarching hypothesis, inadvertently dominates the outcome of the risk analysis, and facilitates model development and weight assignment; 3) sub-hypotheses represent models with parameters that cannot be reliably estimated within the assessment model and are therefore fixed in the models; 4) multiple metrics are used to evaluate the reliability of the models and the plausibility of the hypotheses they represent; 5) model fit only plays a limited role in metrics used to evaluate models; 6) an efficient approach to eliminate unlikely hypotheses. This approach was illustrated by applying it to the stock assessment of EPO bigeye tuna and is the first attempt at a comprehensive risk analysis to evaluate the harvest control rule for tropical tunas in the EPO in a probabilistic framework that considers multiple alternative models. Some improvements could be considered in future developments of this approach. For example, research will involve fuller estimation of uncertain parameters (*e.g. M*), inclusion of additional data or data-based priors

(*e.g.* information on growth from growth increment data), and identification and removal of model misspecification. Hopefully, this will reduce the number of models needed to represent the alternative hypotheses. Further investigation will be conducted of the use of MCMC algorithms to improve the probability distributions, particularly at the tails of the distributions.

## 5.2. General

### 5.2.1. Number of hypotheses

Evaluating multiple alternative hypotheses about states of nature has become common practice in fisheries stock assessment and management. However, the number of possible hypotheses, and models representing those hypotheses, can easily become impractical for contemporary stock assessment models that are computationally intensive. For example, in the bigeye tuna application, models were developed by combining the Level 2B hypotheses (*misspecified growth*, *dome-shape selectivity*, or *increased natural mortality*) with the Level 2A hypotheses (Figure 2). Fortunately, several of the Level 2A hypotheses-based models were eliminated for a variety of reasons and not applied to all Level 2B hypotheses (Figures 2 and 3). An alternative approach would have been to evaluate many models, perhaps weighted by their fit to the data, without considering the relative reliability among models (*e.g.* applying diagnostics) in detail, if at all. The parameter values in such an approach are often selected by applying a grid to the space of all possible values of model parameters and can therefore involve thousands of models. However, we consider that doing so is inappropriate for the following reasons: 1) there is less guarantee that a model has converged on the global minimum of the objective function when only model fit is evaluated; 2) equal weighting of hypotheses ignores *a priori* any ancillary information about the plausibility of each hypothesis, 3) some combinations of hypotheses are not plausible; 4) traditional measures of model fit to data are usually biased in complex stock assessment models; and 5) diagnostics are needed to determine if a model is reliable.

Our approach of efficiently eliminating models was based on defining and running "Base" models, each representing a Level 2A hypothesis (Figure 2). If a Base model was eliminated (for example due to the model not meeting convergence criteria), then any models dependent on it (e.g., models at Level 2B or at Level 3) were also eliminated (Figure 3). We consider that this is reasonable in the Level 3 steepness case since for models with ample composition data, the value of steepness generally does not influence model results and only substantially influences reference points. We are less confident for the Level 2B hypotheses (*Growth*, *Selectivity*, and *Natural Mortality*) since these are more likely to have an impact on model performance. The following Base models were eliminated by this procedure (see Figures 2 and 3): the model using a Ricker type stock recruitment curve because did not converge; the *Unrepresentative longline length composition* model because it had too much "doming" in the "empirical" selectivity for the longline fishery with asymptotic selectivity; and, the *Index not representative* model because it did not improve the regime shift. It is unclear if any of these models would be improved by estimating parameters that would both fix the recruitment regime shift and the misfit to the length composition data. Future research should investigate these models further.

### 5.2.2. Steepness

Steepness of the Beverton-Holt stock-recruitment relationship is notoriously difficult to estimate in fisheries stock assessment models (*e.g.* Lee et al. 2012). Estimates of steepness are frequently biased and often estimated at the parameter estimation bounds (1.0 or 0.2). The estimates are highly influenced by potential regime shifts in recruitment. Therefore, they are considered unreliable in may stock assessments. It is common to use priors from meta-analysis to inform estimation of the value of steepness in stock assessment models, but these meta-analyses are based on estimates of steepness that are subject to the same biases and therefore may also be unreliable. Unfortunately, it may therefore be inappropriate

to use the information about steepness contained in the data or data-based priors from meta-analysis, particularly for tuna, which are highly fecund pelagic spawners that exhibit large recruitment variability. We addressed this issue by constructing several sub-hypotheses representing discrete values of steepness on the lowest level of the hierarchy and did not allow the fit to the data to inform these sub-hypotheses. Ideally, the reason for the bias in the estimates of steepness would be identified and removed so that steepness can be treated like the other parameters in the model, but we leave this to further research and instead recommend our sub-hypothesis approach.

The weightings assignments were all based on the model with recruitment independent of stock size (steepness = 1.0). Despite the fact that we considered this reasonable, as mentioned above, it is possible that some weightings may have differed based on varying levels of steepness (e.g. the recruitment regime shift metric), but we consider that this is a minor impact compared to the bias caused by allowing the model fit to influence steepness and the computational burden of evaluating all the diagnostics *etc*. for each value of steepness.

#### 5.2.3. Within-branch probabilities

We have used a conditional probability-type approach to rescale weights assigned to each model and to combine models to construct the probability distributions for the quantities of interest (e.g. $F_{cur}/F_{MSY}$). This involved rescaling weights to hypotheses on the same level of the hierarchy in the same branch (*e.g.* all models that represent the regime shift in recruitment being real (Figure 3): *Environment* and *Ricker* models), without regard to models and hypotheses on other branches of the hierarchy. This approach avoids the concern that the resulting probability for an overarching hypothesis is overly influenced by the number of hypotheses (models) that represent it. However, the approach puts more emphasis on the weights assigned to the overarching hypotheses. For example, in the bigeye tuna application, there would always be 20% weight given to the regime shift is real branch of the hierarchy, no matter what weights were given to each of the model models on Level 2. This weight is strictly a result of the weights given to the overarching hypothesis by the experts and is not related to the expert judgment about the individual hypotheses, the model fit to the data, or the reliability of the model based, for example, on diagnostics. Therefore, the conditional probability approach might give too much weight to a lower-level model that is considered unreliable, if the higher-level hypothesis is given substantial weight. In this case, it may be reasonable to re-examine the weights given to the overarching hypotheses once the weights for the individual hypotheses have been assigned. However, this is somewhat circular, and it is not clear if it is appropriate. On the other hand, given the issues with using the model fit to weight hypotheses, and overweighting based on the number of hypotheses representing an overarching hypothesis, simply using model fit to assign weights may bias the result towards a single model, too.

#### 5.2.4. Nested hypothesis

The complexity of fisheries and fish stocks, and our limited understanding of their dynamics and structure, often results in having several nested hypotheses about the state of nature, many of these with no data to differentiate them. In Level 2 of our hierarchy (Figure 2) we have explicitly accounted for nested hypotheses within the sub-levels. This nesting helps in formulating the models, assigning weights, and efficiently reducing the number of models as they are eliminated. However, in a different application, nested hypotheses might also be desirable in Levels 1 and 3. Therefore, the approach may have to be modified depending on the application, the issues with the particular assessment, and the data available. Adding nested overarching hypotheses in Level 1 would result in a conditional probability tree based on expert opinion. Nesting of hypotheses on Level 3 would be more case specific.

### 5.2.5. Subjectivity

Although the proposed method produced probability distributions for models and quantities of interest, several aspects of the approach were subjective and qualitative. First, the set of hypotheses, and models that represent them, that were included for consideration in the risk analysis was chosen based on the experts' subjective opinion. Next, many of the weighting metrics were subjective, some more so than others. To reduce the impact of subjectivity on the weighting process, we used discrete weight categories, rather than values on a continuous scale, to assign values to the various weighting metrics. However, even the choices of quantitative values given to each weight category, and the number of weight categories, were subjective. There were four weight categories used: *None* (value = 0), Low (value = 0.25), *Medium* (value = 0.5) and *High* (value = 1). Different values could have been used and would have changed the resulting distributions. In addition, for models considered less plausible, the choices could only be *None* (zero weight) and *Low* (0.25), with no option to select a weight category with a very low weight. An additional category could be added such as *Very Low* (*e.g.* 0.1 weight) to address this. However, research is needed to make the metrics more objective and quantitative.

Care needs to be taken when assigning weights. Historically, subjectivity entered into this process whenever precautionary principles played a role in creating the stock assessment model because often this would result in the adoption of conservative options for model parameters (*e.g.* choosing a lower value of steepness). However, this is not consistent with the risk analysis approach. The risk assessment and other management advice should be based on the best available information. When this management advice is presented to managers, any uncertainty about the information on which this advice is based also should be presented. This allows the managers to take uncertainty into consideration when forming management actions. Therefore, the weighting should be based on the best available information.

### 5.2.6. Data weighting

In the idealized approach (Section 2.1), the data and appropriate priors inform the model structure and parameter values. Unfortunately, due to the complex nature of stock assessment models, our lack of complete understanding about the system, and the inability of parameter estimation methods to deal with large numbers of correlated parameters, stock assessment models are not well constructed in a statistical sense. The parameter estimates are based on large amounts of data, are often overly precise, and biased by model misspecification. Standard statistical measures typically overwhelmingly support a single hypothesis compared to the alternatives, while even the best model violates many of the statistical assumptions. Therefore, simply using a measure of fit (*e.g.* AIC) to weight models is inappropriate. We have addressed this by using a variety of metrics to weight the models, and we have evaluated model fit based on the range of AIC values seen among the models rather than on standard recommendations. We also avoided biases in estimates of the steepness of the Beverton-Holt stock-recruitment relationship by using a discrete range of values as sub-hypotheses to represent models with different values of steepness. However, this approach conflicts with our procedure for creating probability distributions of the management quantities conditional on the model. These probability distributions are created using the standard deviation (standard error) of those quantities obtained from the estimation procedure, which only represents parameter estimation error under the assumptions of the particular model, and are typically biased if those assumptions are wrong. That is, for model structure uncertainty we ignore standard statistical procedures, but for parameter uncertainty we use them. This conflict could in theory be improved by removing model misspecification, modelling parameter temporal variation, estimating parameters that are assumed known, and estimating the appropriate variances for the appropriate likelihood functions, including taking correlation among residuals into consideration. However, in practice it is unlikely that all the conflicts will be addressed for a particular application. Furthermore, the models may not converge on the global minimum of the objective function, and simplification of the model will

be required to address this problem. Therefore, appropriate statistical weighting for hypothesis testing and representing uncertainty will still need to be considered, and further research is needed.

### 5.2.7. Tails

The most important probability evaluations in the IATTC tropical tuna HCR are related to the tails of the probability distribution of the quantity of interest [*e.g.* $P(S_{cur} < S_{LIMIT}) > 10\%$]. Therefore, to have an accurate evaluation of the HCR, the tails of the probability distribution must be well estimated. The approximations we use here are based on the normal distribution, which is a symmetrical probability distribution, and thus, our estimates of tail probabilities may be biased if the true distribution is asymmetrical, particularly when transformations were used to calculate the standard deviation. The estimated standard deviation (standard error) could also be biased and influence the tails of the distribution. We compared the normal approximation to the posterior distribution obtained from MCMC and found that although the posterior distribution and the normal approximation were centered on similar values, the tails of the posterior distributions were fatter, and some posterior distributions were not symmetrical, all of which may affect to some degree the appropriateness of the probabilities calculated assuming a normal distribution (see Figure 8). If probabilities in the tail of the distribution are continued to be used, more research is needed to more accurately represent the tails, including use of MCMC methods.

### 5.2.8. Convergence

Ensuring convergence of the estimation routine is an important part of model development and parameter estimation. Achieving convergence may get more challenging as the number of parameters and the amount and types of data increase. Contemporary statistical integrated fisheries stock assessment models are highly parameterized and use several different types of data, and convergence issues are common. To address this problem, our approach includes a weighting metric based on metrics of model convergence. In the bigeye tuna application, we simply used the metric that the estimated Hessian matrix had to be positive definite to be included in the risk analysis (Section 2.2.2.b). Although a non-positive definite Hessian generally means that that model did not converge, and the parameters are not reliable, it does not definitively mean that the model does not correctly represent the hypothesis or that the hypothesis has no support from the data. It may mean simply that the initial parameter values caused the model to wander into unreasonable parameter estimation space or that there is not enough information to estimate all the model parameters specific to the hypothesis. Therefore, substantial effort should be made to achieve convergence for models not producing a positive definite Hessian matrix. Of course, even if the Hessian is positive definite, that does not guarantee that the estimation procedure has reached the solution corresponding to the global minimum. There may be solutions in a different region of the parameter space that represent very different management advice that could have similar or even better objective function values. Multiple local minima of similar likelihood values suggest that the uncertainty represented based solely on parameter uncertainty is an underestimation for that model and some attempt should be made to represent the full uncertainty (e.g. using MCMC methods that better describe the parameter space). This is a common problem with complex models, becoming even more problematic when many models are considered since there is less time available to fully evaluate each model. This problem is compounded by the fact that when multiple models are used, there is more of a tendency to add additional parameters to represent a wider range of hypotheses. We did not use the maximum gradient criterion to evaluate whether the model converged since it is not clear what the criterion should be; we leave that to future research.

### 5.2.9. Residuals

Evaluating residuals of model fits to data is standard practice in statistical modelling. However, this becomes more difficult as the models become more complex and multiple data types are used. The bigeye tuna example considered three types of residuals typically used to evaluate the performance of fisheries stock assessment models (Section 2.2.2b): 1) residuals of the fit to indices of abundance, 2) residuals of the fit to composition data, and 3) recruitment deviations. However, we decided to ignore the residuals when determining weights for the models because, in most cases, the residuals were similar for all models, and other weighting factors were considered more important. In addition, the composition residuals were also covered by the *W(Empirical selectivity)* weighting metric. A more quantitative approach is needed to evaluate residuals in general, and specifically for weighting the models. The criteria need to address any violation of the likelihood function assumptions, including the shape of the distribution of the residuals (including outliers), the magnitude of the residuals, and lack of independence. In stock assessment models this evaluation can be complicated because the likelihood function measures the total misfit to the data and therefore not only represents the sampling distribution, but also model misspecification and unmodelled process variation, which are common in stock assessment models. Approaches have been developed to estimate the variance parameter of the likelihood function and this may remove the need to evaluate the magnitude of the residuals.

We introduced the *W(Empirical selectivity)* diagnostic to compare the selectivity implied by the model estimates with the selectivity pattern assumed in the stock assessment. The "empirical" selectivity is dependent on the model, but is not constrained by the assumed selectivity curve, and hence the use of the terminology "empirical". The "empirical" selectivity is calculated as the model estimated catch-at-age (or length) divided by the estimates of the population numbers-at-age (or length). This diagnostic provides similar information to measures of the overall fit of the observed and predicted composition data (plotted for a fishery averaged over all years, perhaps weighted by the sample size), but gives more emphasis to older and larger fish. This diagnostic was very informative in the bigeye assessment when assessing model fit to the length composition for large fish in the longline fishery that had asymptotic selectivity. We recommend that calculating this diagnostic be standard practice in fisheries stock assessment. Consideration should be given to weighting the calculations by the composition sample size when taking the average and plotting the diagnostic annually.

## 5.3. Bigeye application

### 5.3.1. Improving the assessment

The goal of the risk analysis for EPO bigeye tuna was to determine the probability of exceeding the target and limit reference points under current and alternative management actions, such as fishery closures. The risk assessment for bigeye tuna was hierarchically structured around two major issues with the stock assessment: 1) the estimated regime shift in recruitment and 2) the misfit to the length composition data for the longline fishery with assumed asymptotic selectivity. Some of the hypotheses used in the risk analysis ascribe issue (1) to a modelling artifact, while others postulate that recruitment really did increase. Results to date suggest that issue (2) could be due to model misspecification on growth, natural mortality and other processes. Preferably, it would be best to resolve these issues rather than include them in a risk analysis, therefore research on these issues should continue. In addition to these issues, the assessment results had become highly sensitive to new data points in the indices of relative abundance derived from the longline fishery. The over-sensitivity of the previous assessment to the addition of new data, which was the reason for not using it for management advice during the two years prior to this work, is no longer an issue for most of the models according to the retrospective analysis. The previous assessment sensitivity was due to the addition of a new year (four quarters) of abundance index data

(SAC-09 INF-B), while the retrospective analysis tests the removal of all data one year at a time, and therefore although the comparison is not perfect it suggest that this issue no longer remains. The new spatio-temporal modelling framework used to standardize the indices of abundance may have removed this sensitivity. The risk analysis for bigeye tuna has identified two major avenues of research to improve the stock assessment: 1) estimating growth and 2) investigating the differences between the models of different time spans. The models that estimate growth have a combined 58% of the model weight and therefore obtaining data to improve the estimates of growth (particularly for the older/larger fish) should be a priority. The bimodal probability distributions for the current status relative to reference points is driven by several factors, (e.g. growth, natural mortality, selectivities), but mainly by the time span of the models. This analysis used two classes of model: short-term and medium- term. Research should be conducted using models of different time spans and to investigate why short-term models estimates equilibrium catches much higher than the catch in the years prior to the start of the model.

#### 5.3.2. Model time-spans

There are several research avenues that could be pursued to investigate differences between the models of different time span. One such research would be to start with medium-term models and drop data sets to see which causes different results. Another would be to fit short-term models to an estimate of the equilibrium catch (e.g. the average of the 5 years previous to the start of the model), although this would be similar to the medium-term model only using catch prior to the start of the short-term model. Further research could be conducted using historical time span models starting when the earliest commercial catches are available.

#### 5.3.3. Growth

Model results are highly influenced by how growth is parameterized; therefore it is important to use the available growth data properly and evaluate any potential improvements in both data and estimation methods. The models using fixed growth parameters are based on an external analysis that integrated both the otolith age-length data and the individual growth-increment from tagging data. However, only the otolith age-length data is used when growth is estimated inside the stock assessment model. The tagging growth-increment data should be included in the assessment model, but this functionality is not currently available in Stock Synthesis. An alternative approach is to use the external analysis estimates and their covariance matrix to construct prior distributions for use in the stock assessment model. Stock Synthesis does not currently have the capability to include joint priors so they would have to be included as independent priors on each parameter, which may create issues for example for highly correlated growth parameters. Estimation of growth inside the stock assessment model has several advantages: 1) information on growth is also contained in other types of data (such as length composition data) in an integrated model; 2) length- or age-based selectivity is automatically taken into account; and 3) length based sampling of age-length data can be addressed using age conditioned on length composition data. Therefore, the most appropriate and priority approach is to integrate the tagging growth increment data into the stock assessment model.

#### 5.3.4. Natural mortality

Natural mortality (*M*) is one of the most uncertain and influential parameters in fisheries stock assessment models. We included two hypotheses about natural mortality: *Adult M* that estimated the natural mortality of adults and *Movement* that estimated the natural mortality of pre-adults and assumed that it was the same for adults as well. Both hypotheses kept the same ratio as that of the models not estimating *M* for the difference between male and female natural mortality. Differences between the assumed fixed *M* and those estimated by the additional hypotheses could represent either incorrectly assumed natural mortality or unmodeled movement (*M* as a proxy for movement). Other hypotheses about natural

mortality could also be included in the analysis. For example, the Lorenzen curve was used in a previous analysis (Valero et al. 2019) and it also removed the regime shift in recruitment. Other analyses not presented here have shown that natural mortality for juveniles does not reduce the estimated recruitment regime shift (Document SARM-9-INF-B), but only estimating natural mortality of pre-adults (and not extending it to adults) does. The natural mortality based on the Lorenzen curve also increases the natural mortality of pre-adults and it is likely that the increase of natural mortality for these ages using the Lorenzen curve is what reduces the recruitment regime shift. This is because the increase of catches in the floating object fisheries (catching primarily pre-adults) would have a smaller effect on indices of adult abundance because less fish would reach the longline fishery since they had moved out of the EPO or died. The bigeye assessment assumes that the differences in sex ratios with size is due to differences in natural mortality, but due to the limited information of sex specific growth, particularly for older ages, there may be an effect of sex-specific growth. Further research is needed to determine the appropriate levels of length-, age- and sex-specific bigeye tuna natural mortality.

#### 5.3.5. Improving the risk assessment for bigeye tuna

A better way to eliminate models or evaluate more models is needed. Due to the large number of models (48) and the computational demands of some of the diagnostics (*e.g.* ASPM, $R_0$ profile, retrospective analysis, catch curve diagnostic) it was not feasible to conduct all the diagnostics on the models with the different values for steepness ($h$) or some of the Level 2B hypotheses in combination with every Level 2A hypothesis. Therefore, we eliminated models that performed poorly with $h$ = 1.0 and assumed that the weightings for the models that had $h$ < 1.0 were the same as those for $h$ = 1.0, except the *W(convergence)* weight. Since the diagnostics were not run and evaluated, we do not know if this assumption is correct. We found that the *W(Fix regime)* weight scores were different among some of the steepness values (not used it in the model weighting). Making the diagnostics quantitative and automating them will allow for more models to be evaluated.

The risk analyses presented here relied on quantities of interest (such as ratios of stock status relative to reference points) for which approximations were made on the type of distribution as well the standard error describing its uncertainty. The validity or potential departure of this approximations should be tested for example via simulation work or MCMC techniques. The extent to which the model weighting scheme could be treated as probabilities should also be further evaluated.

#### 5.3.6. Management implications

The inconsistency of management implications between the expected value and the probability statements is due to the asymmetric nature of the combined probability distribution. The current closure produces a 50% probability that $F_{cur} < F_{MSY}$. However, the expected value (or model weighted MLEs) of $F_{cur}/F_{MSY}$ suggests that the current fishing mortality is 7% too high. This is caused by the asymmetric nature of the combined probability distribution. Even though the distributions of the individual model quantities of interest are assumed to be normal, when they are combined with different model weights, the combined distribution becomes asymmetrical. The expected value is the mean and the probability-based statement is related to the median. However, this difference is relatively minor, and has smaller management implications than the bimodality obtained for all probability distributions. Results from the bigeye risk analysis essentially fall in between two possible states (an optimistic and a pessimistic, relative to reference points) that cannot be discerned based on data, model valuation or other criteria currently available.

The use of a limit reference point based on equilibrium $S_0$ causes a dilemma, particularly when evaluating models with different time spans. Equilibrium $S_0$ is a function of $R_0$, which itself is a function of the average recruitment over the modelling time frame and assumptions about model initial conditions, adjusted

appropriately by the stock-recruitment relationship. The calculation of $R_0$ is typically defined over the time span of the model, which is convenient because no decision on the years used to calculate $R_0$ needs to be made. However, when models with different time spans are included in the risk analysis, they are no longer consistent, and the evaluation of this reference point may be more a function of the time span of the model and not the hypothesis being represented by that model. In the bigeye tuna application, the short-term models assume that the initial low recruitment is unrepresentative and should not be use in the calculation of $R_0$. Therefore, the results may be appropriate. In other applications, it may be appropriate to choose a range of common years across models to estimate $R_0$ that differ from the time span of the model to make the results consistent among models. However, this concept does bring in to question the construction of limit reference points when there may be regime shifts in recruitment and the definition of years to calculate $R_0$ should be given more explicit consideration, particularly when potential changes in productivity may be confounded with fishery impacts on the stock. Dynamic reference points like $S_{MSY\_d}$ do not use average recruitment and are therefore not dependent on defining the time period.

As stated before, all the probability distributions for the management quantities for bigeye tuna show two modes. This produces a dilemma because management action based on fishing at $F_{MSY}$ should not simply take the average value of two different states of nature, since it is unclear which state of nature is correct and this will either highly under or over exploit the stock. A precautionary approach might imply basing management on the pessimistic models, but this might severely under exploit the stock and would require a substantial reduction in the fishing effort. One approach might be to project the stock into the future assuming a high biomass while using the $F_{MSY}$ estimated using the low biomass and vice versa, to determine the impacts and potential tradeoffs of taking management action assuming the wrong state of nature.

## 6. CONCLUSIONS

Given these results, the recommended way forward would be to put substantial effort into collecting data, improving the stock assessment models, identifying and correcting model misspecifications, particularly as it relates to the two modes in the probability distributions for the management quantities, and evaluating management strategies robust to uncertainty. We also should acknowledge that there may always be unresolved issues in knowledge, their impact on taking appropriate management action, and the inherent limits of modelling complex and changing natural systems and their fisheries. An alternative, or complementary approach while both data and modelling approaches improve, would be to conduct MSE to evaluate setting management actions based on simpler models or trends in data such as empirical HCRs. The models and their weighting developed here could be used as a basis for developing MSE operating models.

Given the substantial uncertainty in stock assessments in general, management decisions should not simply be based on point estimates from a single base case model or even point estimates derived from an average from multiple models. Management should take into consideration the uncertainty in the estimates, model structure and other components of the system (implementation, *etc.*). Developing management strategies that incorporate, and are evaluated via MSE to be robust to, the different forms of unavoidable uncertainties involved in fishery management are a formal way to evaluate management actions designed to achieve fisheries objectives.

**REFERENCES**


Aires-da-Silva, A., M. N. Maunder, and P. K. Tomlinson. 2010. An investigation of the trend in the estimated recruitment for bigeye tuna in the eastern Pacific Ocean. Document BET-01-06. Inter. Amer. Trop. Tuna Comm. Document BET-01-02b. External review of IATTC bigeye tuna assessment. La Jolla, California, USA. 3-7 May 2010.

Burnham, K.P., and Anderson, D.R. 1998. Model selection and multimodel inference: a practical information-theoretic approach. 1st ed. Springer, New York.

Butterworth, D.S. 2007. Why a management procedure approach? Some positives and negatives. ICES Journal of Marine Science 64, 613–617.

Deriso, R.B., 1980. Harvesting strategies and parameter estimation for an age-structured model. Can. J. Fish. Aquat. Sci. 37, 268–282.

IATTC 2013. Ecosystem Considerations. IATTC Document SAC-04-08. http://www.iattc.org/Meetings/Meetings2013/SAC-04/Docs/_English/SAC-04-08_Ecosystem%20considerations%20Ecological%20and%20Physical%20changes%20in%20the%20EPO.pdf

IATTC. 2016. Harvest Control Rules for tropical tunas (yellowfin, bigeye, and skipjack). Resolution C-16-02. 4pp.

Maunder, M. N. and R.B. Deriso. 2014. Proposal for biomass and fishing mortality limit reference points based on reduction in recruitment.

Nakatsuka, S. 2017. Management strategy evaluation in regional fisheries management organizations − How to promote robust fisheries management in international settings. Fisheries Research 187: 127–138.

Punt, A.E., Hilborn, R. 1997. Fisheries stock assessment and decision analysis: the Bayesian approach. Reviews in Fish Biology and Fisheries 7, 35–63

Punt, A.E., Butterworth, D.S., de Moor, C.L., De Oliveira, J.A.A., Haddon, M. 2016. Management strategy evaluation: best practices. Fish and Fish. 17: 303-334.

Schnute, J.T., and R. Hilborn. 1993. Analysis of contradictory data sources in fish stock assessment. Can. Fish. Aquat. Sci. 58: 191 6-1 923.

Valero, J. L., Aires-da-Silva, A. and Maunder, M. N. 2018. Exploratory spatial stock assessment of Bigeye tuna (*Thunnus obesus*) in the EPO. Inter-Amer. Trop. Tuna Comm., 9th Scient. Adv. Com. Meeting. SAC-09.

Valero, J. L., Maunder, M., Xu, H., Minte-Vera, C. V., Lennert-Cody, C., Aires-da-Silva, A. 2019. Spatial stock assessment model options for bigeye tuna (*Thunnus obesus*) in the EPO and beyond. Review of the stock assessment of bigeye tuna in the eastern Pacific Ocean. La Jolla, California (USA), 11-15 March 2019. WSBET-02-09.

Valero, J. L., Maunder, M., Xu, H., Minte-Vera, C. V., Lennert-Cody, C., Aires-da-Silva, A. 2019. Investigating potential causes of misspecification-induced regime shift in recruitment in the EPO bigeye tuna (*Thunnus obesus*) assessment. Review of the stock assessment of bigeye tuna in the eastern Pacific Ocean. La Jolla, California (USA), 11-15 March 2019. WSBET-02-08.

Valero, J. L., Maunder, M., Xu, H., Minte-Vera, C. V., Lennert-Cody, C., Aires-da-Silva, A. 2019. Summary of modeling work on evaluating bigeye tuna recruitment shift hypotheses. Inter-Amer. Trop. Tuna Comm., 10th Scient. Adv. Com. Meeting. SAC-10 INF-G.

Valero, J.L. and Aires-da-Silva, A. 2020. 1st Workshop on Management Strategy Evaluation (MSE) for Tropical Tunas: Overview, objectives and performance metrics. La Jolla, California (USA), 9-10 December 2019.

**TABLE 1.** Weighting criteria for the recruitment regime shift metric.

| Regime Shift category | Weight category | Weight value |
|---|---|---|
| $1.75 < R_{shift}$ | None | 0 |
| $1.50 < R_{shift} \leq 1.75$ | Low | 0.25 |
| $1.25 < R_{shift} \leq 1.5$ | Medium | 0.5 |
| $R_{shift} \leq 1.25$ | High | 1 |
| Regime shift is real | NA | 1 |

**Table 2** Model names and acronyms.

| | |
|---|---|
| Env-Fix | Environment, Fixed |
| Env-Gro | Environment, Estimate growth |
| Env-Sel | Environment, Dome selectivity |
| Env-Mrt | Environment, Adult mortality |
| Rcr | Ricker |
| Ind | Index not representative |
| Srt-Fix | Short-term, Fixed |
| Srt-Gro | Short-term, Estimate growth |
| Srt-Sel | Short-term, Dome selectivity |
| Srt-Mrt | Short-term, Adult mortality |
| Mov | Pre-adult movement |
| Gro | Estimate growth |
| Sel | Dome selectivity |
| Mrt | Adult mortality |
| Cmp | Unreliable longline composition |

**TABLE 3.** *W(Expert)* weights assigned by each expert to the alternative models.

| | Level 2A | | | | | | Level 2B | | | | | | Probability |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Env-Fix | High | High | High | High | High | High | Low | Low | Medium | Low | Low | Medium | 0.13 |
| Env-Gro | High | High | High | High | High | High | High | High | Medium | High | High | High | 0.33 |
| Env-Sel | High | High | High | High | High | High | High | High | Medium | High | Medium | Medium | 0.27 |
| Env-Mrt | High | High | High | High | High | High | Medium | Medium | Medium | Medium | Medium | Medium | 0.18 |
| Rcr | Low | Low | Low | Low | Low | Low | NA | NA | NA | NA | NA | NA | 0.09 |
| Ind | Low | Medium | Medium | Low | Low | Low | NA | NA | NA | NA | NA | NA | 0.08 |
| Srt-Fix | Medium | High | Low | High | High | Medium | Medium | Low | High | Low | Low | High | 0.07 |
| Srt-Gro | Medium | High | Low | High | High | Medium | Medium | Medium | High | High | High | High | 0.12 |
| Srt-Sel | Medium | High | Low | High | High | Medium | Medium | Medium | Low | Medium | Medium | Low | 0.07 |
| Srt-Mrt | Medium | High | Low | High | High | Medium | Medium | Medium | Low | Medium | Medium | Medium | 0.07 |
| Mov | High | High | Medium | High | High | High | Medium | High | Medium | Low | Medium | Medium | 0.11 |
| Gro | High | Medium | Medium | High | High | High | High | High | Medium | High | High | High | 0.17 |
| Sel | High | High | Medium | Medium | High | Medium | High | High | Medium | High | Medium | Medium | 0.13 |
| Mrt | Medium | Medium | Medium | Medium | Medium | High | Medium | Medium | Medium | Medium | Medium | Medium | 0.07 |
| Cmp | Medium | Medium | Medium | Medium | Medium | Medium | NA | NA | NA | NA | NA | NA | 0.12 |

**TABLE 4.** Maximum gradient components for the alternative models. NPD indicates that the hessian was not positive definite. The *Ricker* model did not have a steepness parameter.

| | *h* = 1.0 | *h* = 0.9 | *h* = 0.8 | *h* = 0.7 |
|---|---|---|---|---|
| Env-Fix | 0.0002 | NPD | NPD | NPD |
| Env-Gro | 0.00009 | 0.00003 | 0.00008 | 0.00002 |
| Env-Sel | 0.00007 | 0.00006 | 0.00007 | 0.00005 |
| Env-Mrt | 0.00003 | 0.0002 | 0.00005 | 0.00007 |
| Rcr | NPD | NA | NA | NA |
| Ind | 0.00005 | NPD | NPD | NPD |
| Srt-Fix | 0.00005 | 0.00007 | 0.006 | NPD |
| Srt-Gro | 0.00008 | 0.0003 | 0.001 | 0.002 |
| Srt-Sel | 0.00007 | 0.00002 | 0.00004 | 0.00001 |
| Srt-Mrt | 0.00008 | 0.00005 | 0.01 | 0.003 |
| Mov | 0.00007 | 0.0001 | 0.0001 | 0.0002 |
| Gro | 0.00003 | 0.00008 | 0.00005 | 0.0001 |
| Sel | 0.001 | 0.0009 | 0.001 | 0.0002 |
| Mrt | 0.00006 | 0.00006 | 0.001 | 0.00007 |
| Cmp | 0.00003 | 0.0001 | 0.0003 | 0.0003 |

**TABLE 5.** AIC values calculated using all the data except the otolith age-length data, the score based on the formula and the associated recalled probability.

| | *h* = 1.0 | *h* = 0.9 | *h* = 0.8 | *h* = 0.7 | Score (*h* = 1.0) | Probability |
|---|---|---|---|---|---|---|
| Env-Fix | 5245.92 | NA | NA | NA | 0.28 | 0.12 |
| Env-Gro | 5125.20 | 5125.16 | 5125.44 | 5125.29 | 1.00 | 0.42 |
| Env-Sel | 5178.30 | 5178.82 | 5179.56 | 5180.66 | 0.68 | 0.29 |
| Env-Mrt | 5226.02 | 5226.96 | 5228.36 | 5230.50 | 0.40 | 0.17 |
| Rcr | NA | NA | NA | NA | NA | NA |
| Ind | 5568.80 | NA | NA | NA | NA | NA |
| Srt-Fix | 3890.20 | 3891.98 | 3894.98 | NA | 0.25 | 0.06 |
| Srt-Gro | 3858.33 | 3859.15 | 3860.31 | 3862.33 | 1.00 | 0.23 |
| Srt-Sel | 3876.74 | 3877.42 | 3878.42 | 3879.84 | 0.57 | 0.13 |
| Srt-Mrt | 3889.82 | 3890.94 | 3892.82 | 3895.28 | 0.26 | 0.06 |
| Mov | 5251.30 | 5253.16 | 5255.14 | 5258.70 | 0.25 | 0.06 |
| Gro | 5127.69 | 5128.34 | 5129.21 | 5130.38 | 0.99 | 0.23 |
| Sel | 5185.30 | 5186.64 | 5188.32 | 5190.44 | 0.64 | 0.15 |
| Mrt | 5238.62 | 5240.76 | 5243.40 | 5246.70 | 0.33 | 0.08 |
| Cmp | 1928.73 | 1931.69 | 1934.99 | 1938.76 | NA | NA |

**TABLE 6.** *W(Plausible parameter estimates)* weights assigned by each expert to the alternative models, where appropriate.

| Model | | | | | | | Probability |
|---|---|---|---|---|---|---|---|
| Env-Fix | NA | NA | NA | NA | NA | NA | 0.37 |
| Env-Gro | High | Medium | High | High | High | High | 0.34 |
| Env-Sel | Low | Medium | Low | Low | Low | Medium | 0.13 |
| Env-Mrt | High | Low | Medium | Low | Medium | Low | 0.16 |
| Ind | NA | NA | NA | NA | NA | NA | 0.14 |
| Srt-Fix | NA | NA | NA | NA | NA | NA | 0.14 |
| Srt-Gro | High | Medium | Medium | High | High | High | 0.11 |
| Srt-Sel | Medium | Medium | High | Medium | Medium | Medium | 0.08 |
| Srt-Mrt | High | Low | Low | High | High | High | 0.10 |
| Mov | High | Medium | Medium | Medium | High | Medium | 0.09 |
| Gro | High | Medium | High | High | High | High | 0.12 |
| Sel | Low | Medium | Low | Low | Low | Medium | 0.05 |
| Mrt | High | Low | Medium | Low | Low | Low | 0.05 |
| Cmp | NA | NA | NA | NA | NA | NA | 0.14 |

Table 7. Weights for *W(Plausible F)* and *W(Plausible initial catch)*.

| Model | Plausible F | | | | | | Plausible initial conditions | Probability |
|---|---|---|---|---|---|---|---|---|
| Env-Fix | High | High | Medium | High | High | High | Medium | 0.24 |
| Env-Gro | Medium | Medium | High | Medium | Medium | Low | High | 0.24 |
| Env-Sel | Low | Medium | Medium | Low | Medium | Low | High | 0.21 |
| Env-Mrt | Medium | High | High | High | High | High | High | 0.31 |
| Ind | None | None | None | None | None | None | Low | 0.02 |
| Srt-Fix | Medium | High | High | High | High | High | Medium | 0.11 |
| Srt-Gro | Low | High | High | High | Medium | Medium | Medium | 0.09 |
| Srt-Sel | Medium | High | High | High | High | Medium | Medium | 0.10 |
| Srt-Mrt | Medium | High | High | High | High | High | Medium | 0.11 |
| Mov | Medium | High | Medium | High | High | High | High | 0.14 |
| Gro | Low | Low | High | Low | Low | Low | Low | 0.05 |
| Sel | Low | Low | Medium | Low | Low | Low | High | 0.10 |
| Mrt | Medium | High | High | High | Medium | Medium | High | 0.14 |
| Cmp | Medium | Low | High | High | High | Medium | High | 0.13 |

**TABLE 8.** Results of applying the algorithm shown in Figure 1 to calculate the *W($R_0$, ASPM)* component.

| Model | R0 Composition driven | Consistent | ASPM-dev Confidence interval | Consistent | Weight | Probability |
|---|---|---|---|---|---|---|
| Env-Fix | No | NA | Wide | NA | Medium | 0.25 |
| Env-Gro | Yes | Yes | Wide | NA | Medium | 0.25 |
| Env-Sel | Yes | Yes | Wide | NA | Medium | 0.25 |
| Env-Mrt | Yes | Yes | Wide | NA | Medium | 0.25 |
| Ind | NA | NA | NA | NA | High | 0.15 |
| Srt-Fix | Yes | Yes | Wide | NA | Medium | 0.08 |
| Srt-Gro | Yes | Yes | Wide | NA | Medium | 0.08 |
| Srt-Sel | Yes | Yes | Wide | NA | Medium | 0.08 |
| Srt-Mrt | Yes | Yes | Wide | NA | Medium | 0.08 |
| Mov | Yes | No | Narrow | No | Low | 0.04 |
| Gro | Yes | Yes | Narrow | Yes | High | 0.15 |
| Sel | No | NA | Narrow | Yes | High | 0.15 |
| Mrt | Yes | Yes | Narrow | Yes | High | 0.15 |
| Cmp | Yes | Yes | Narrow | No | Low | 0.04 |

**TABLE 9.** Results of the *W(Retrospective)* component.

| Model | | Probability |
|---|---|---|
| Env-Fix | High | 0.40 |
| Env-Gro | Medium | 0.20 |
| Env-Sel | Medium | 0.20 |
| Env-Mrt | Medium | 0.20 |
| Ind | High | 0.11 |
| Srt-Fix | High | 0.11 |
| Srt-Gro | High | 0.11 |
| Srt-Sel | High | 0.11 |
| Srt-Mrt | High | 0.11 |
| Mov | Medium | 0.05 |
| Gro | High | 0.11 |
| Sel | High | 0.11 |
| Mrt | High | 0.11 |
| Cmp | High | 0.11 |

**TABLE 10.** $R_{shift}$ metric (a) and the associated weights (b). The probabilities were only based on the models with $h$ = 1.0.

| Model | $h$ = 1.0 | $h$ = 0.9 | $h$ = 0.8 | $h$ = 0.7 | Probability ($h$ = 1.0) |
|---|---|---|---|---|---|
| Env-Fix | NA | NA | NA | NA | 0.20 |
| Env-Gro | NA | NA | NA | NA | 0.20 |
| Env-Sel | NA | NA | NA | NA | 0.20 |
| Env-Mrt | NA | NA | NA | NA | 0.20 |
| Rcr | NA | NA | NA | NA | 0.20 |
| Ind | None | None | None | None | 0.00 |
| Srt-Fix | NA | NA | NA | NA | 0.14 |
| Srt-Gro | NA | NA | NA | NA | 0.14 |
| Srt-Sel | NA | NA | NA | NA | 0.14 |
| Srt-Mrt | NA | NA | NA | NA | 0.14 |
| Mov | Medium | Medium | High | High | 0.07 |
| Gro | High | High | High | High | 0.14 |
| Sel | Medium | High | High | High | 0.07 |
| Mrt | Medium | Medium | Medium | High | 0.07 |
| Cmp | Medium | Medium | Medium | Medium | 0.07 |

**TABLE 11.** Weights assigned based on the *W(Empirical selectivity)* metric.

| Model | Score | Probability |
|---|---|---|
| Env-Fix | Medium | 0.20 |
| Env-Gro | Medium | 0.20 |
| Env-Sel | High | 0.40 |
| Env-Mrt | Medium | 0.20 |
| Ind | Medium | 0.08 |
| Srt-Fix | Medium | 0.08 |
| Srt-Gro | Medium | 0.08 |
| Srt-Sel | High | 0.17 |
| Srt-Mrt | High | 0.17 |
| Mov | Medium | 0.08 |
| Gro | Medium | 0.08 |
| Sel | High | 0.17 |
| Mrt | Medium | 0.08 |
| Cmp | None | 0.00 |

**TABLE 12.** Weights assigned by each expert to the alternative values of steepness.

| *h* | | | | | | | Probability |
|---|---|---|---|---|---|---|---|
| 0.7 | Low | None | Low | - | None | None | 0.04 |
| 0.8 | Medium | Low | High | - | Low | Low | 0.21 |
| 0.9 | High | Low | High | - | Medium | Medium | 0.31 |
| 1 | Medium | High | Medium | - | High | High | 0.44 |

**TABLE 13.** Model probabilities, expected values of the management quantities, and probabilities of exceeding the reference points.

| | Env-Fix | Env-Gro | Env-Sel | Env-Mrt | Srt-Fix | Srt-Gro | Srt-Sel | Srt-Mrt | Mov | Gro | Sel | Mrt | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P(Model) | 0.01 | 0.13 | 0.05 | 0.02 | 0.04 | 0.22 | 0.11 | 0.07 | 0.01 | 0.24 | 0.09 | 0.02 | 1.00 |
| Fcur/FMSY | 1.82 | 0.82 | 0.99 | 1.25 | 1.84 | 1.42 | 1.36 | 1.57 | 0.81 | 0.59 | 0.73 | 0.89 | 1.07 |
| Fcur/Flimit | 0.96 | 0.47 | 0.58 | 0.69 | 0.97 | 0.78 | 0.77 | 0.84 | 0.47 | 0.34 | 0.43 | 0.50 | 0.60 |
| Scur/SMSY_dyn | 0.34 | 1.32 | 1.02 | 0.69 | 0.32 | 0.56 | 0.59 | 0.45 | 1.31 | 1.85 | 1.53 | 1.16 | 1.09 |
| Scur/Slimit | 0.97 | 3.61 | 2.67 | 2.04 | 0.97 | 1.65 | 1.65 | 1.38 | 3.84 | 5.24 | 4.21 | 3.63 | 3.07 |
| P(Fcur>FMSY) | 1.00 | 0.18 | 0.44 | 0.84 | 1.00 | 0.97 | 0.92 | 0.99 | 0.15 | 0.01 | 0.07 | 0.25 | 0.50 |
| P(Fcur>Flimit) | 0.33 | 0.00 | 0.00 | 0.01 | 0.38 | 0.07 | 0.06 | 0.14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 |
| P(Scur<SMSY) | 1.00 | 0.19 | 0.49 | 0.96 | 1.00 | 1.00 | 1.00 | 1.00 | 0.16 | 0.03 | 0.07 | 0.27 | 0.53 |
| P(Scur<Slimit) | 0.59 | 0.00 | 0.00 | 0.02 | 0.50 | 0.06 | 0.09 | 0.19 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 |

**TABLE 14.** P($F_{cur} > F_{MSY}$) for different closure days adjusted for the *corralito* spatial closure and changes in capacity.

| Closure days | Env-Fix | Env-Gro | Env-Sel | Env-Mrt | Srt-Fix | Srt-Gro | Srt-Sel | Srt-Mrt | Mov | Gro | Sel | Mrt | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.00 | 0.48 | 0.78 | 0.98 | 1.00 | 1.00 | 0.99 | 1.00 | 0.47 | 0.09 | 0.31 | 0.65 | 0.62 |
| 36 | 1.00 | 0.32 | 0.63 | 0.93 | 1.00 | 0.99 | 0.97 | 1.00 | 0.30 | 0.03 | 0.17 | 0.45 | 0.56 |
| 70 | 1.00 | 0.19 | 0.44 | 0.84 | 1.00 | 0.97 | 0.92 | 0.99 | 0.15 | 0.01 | 0.07 | 0.25 | 0.50 |
| 72 | 1.00 | 0.18 | 0.43 | 0.83 | 1.00 | 0.96 | 0.91 | 0.98 | 0.14 | 0.01 | 0.06 | 0.24 | 0.49 |
| 88 | 1.00 | 0.13 | 0.35 | 0.75 | 1.00 | 0.93 | 0.87 | 0.97 | 0.09 | 0.00 | 0.04 | 0.17 | 0.46 |
| 100 | 1.00 | 0.09 | 0.28 | 0.67 | 1.00 | 0.88 | 0.81 | 0.95 | 0.06 | 0.00 | 0.02 | 0.11 | 0.43 |

**TABLE 15.** P($F$cur > $F_{LIMIT}$) for different closure days adjusted for the *corralito* spatial closure and changes in capacity.

| Closure days | Env-Fix | Env-Gro | Env-Sel | Env-Mrt | Srt-Fix | Srt-Gro | Srt-Sel | Srt-Mrt | Mov | Gro | Sel | Mrt | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.97 | 0.00 | 0.04 | 0.17 | 0.89 | 0.39 | 0.37 | 0.57 | 0.00 | 0.00 | 0.00 | 0.00 | 0.21 |
| 36 | 0.79 | 0.00 | 0.01 | 0.06 | 0.67 | 0.19 | 0.18 | 0.33 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 |
| 70 | 0.33 | 0.00 | 0.00 | 0.01 | 0.38 | 0.07 | 0.06 | 0.14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 |
| 72 | 0.30 | 0.00 | 0.00 | 0.01 | 0.36 | 0.06 | 0.06 | 0.13 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 |
| 88 | 0.11 | 0.00 | 0.00 | 0.00 | 0.25 | 0.03 | 0.03 | 0.08 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 |
| 100 | 0.04 | 0.00 | 0.00 | 0.00 | 0.17 | 0.02 | 0.02 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 |

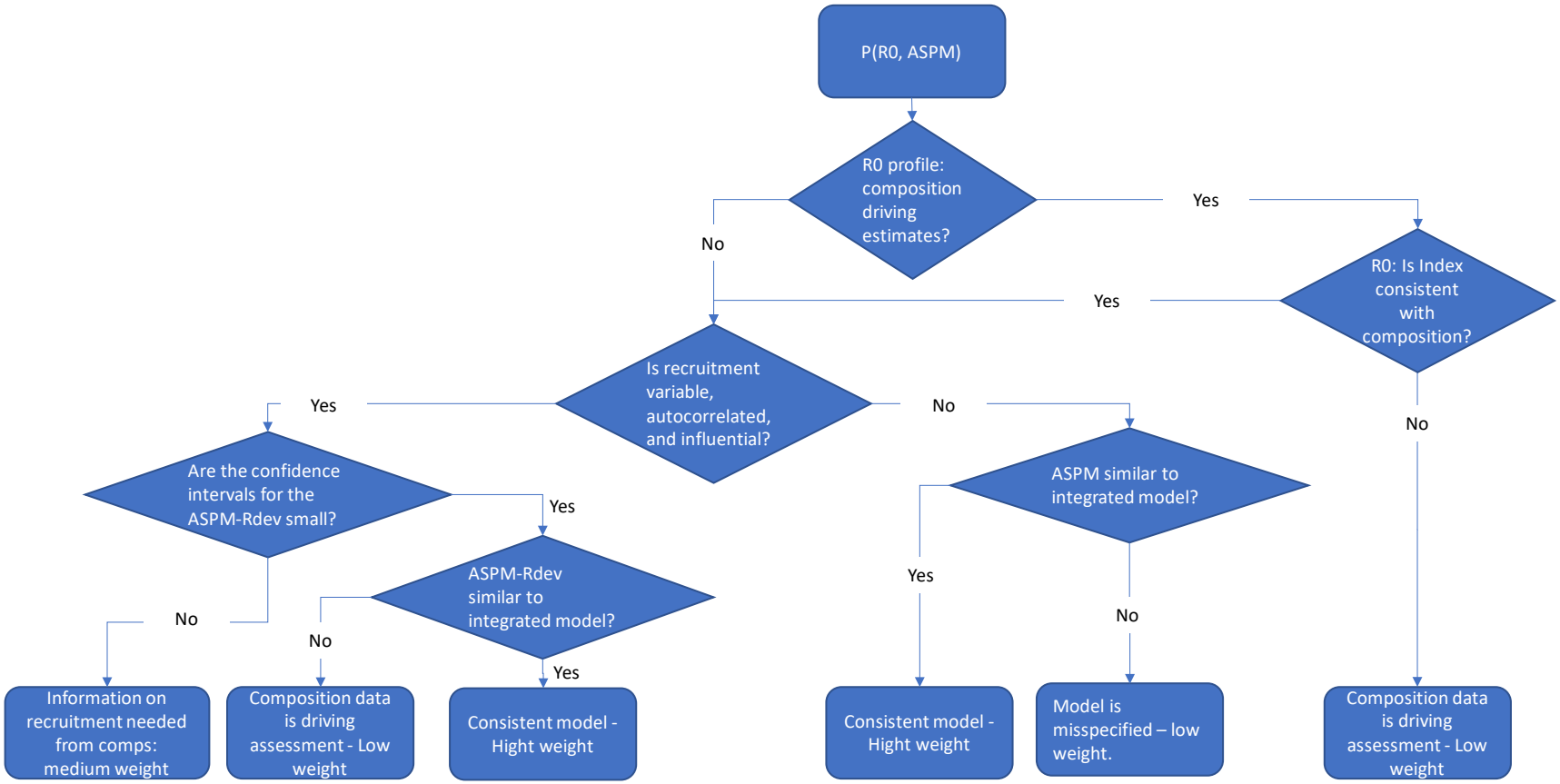


**FIGURE 1**. Algorithm for assigning weights based on the $R_0$ likelihood component profile and ASPM diagnostics.

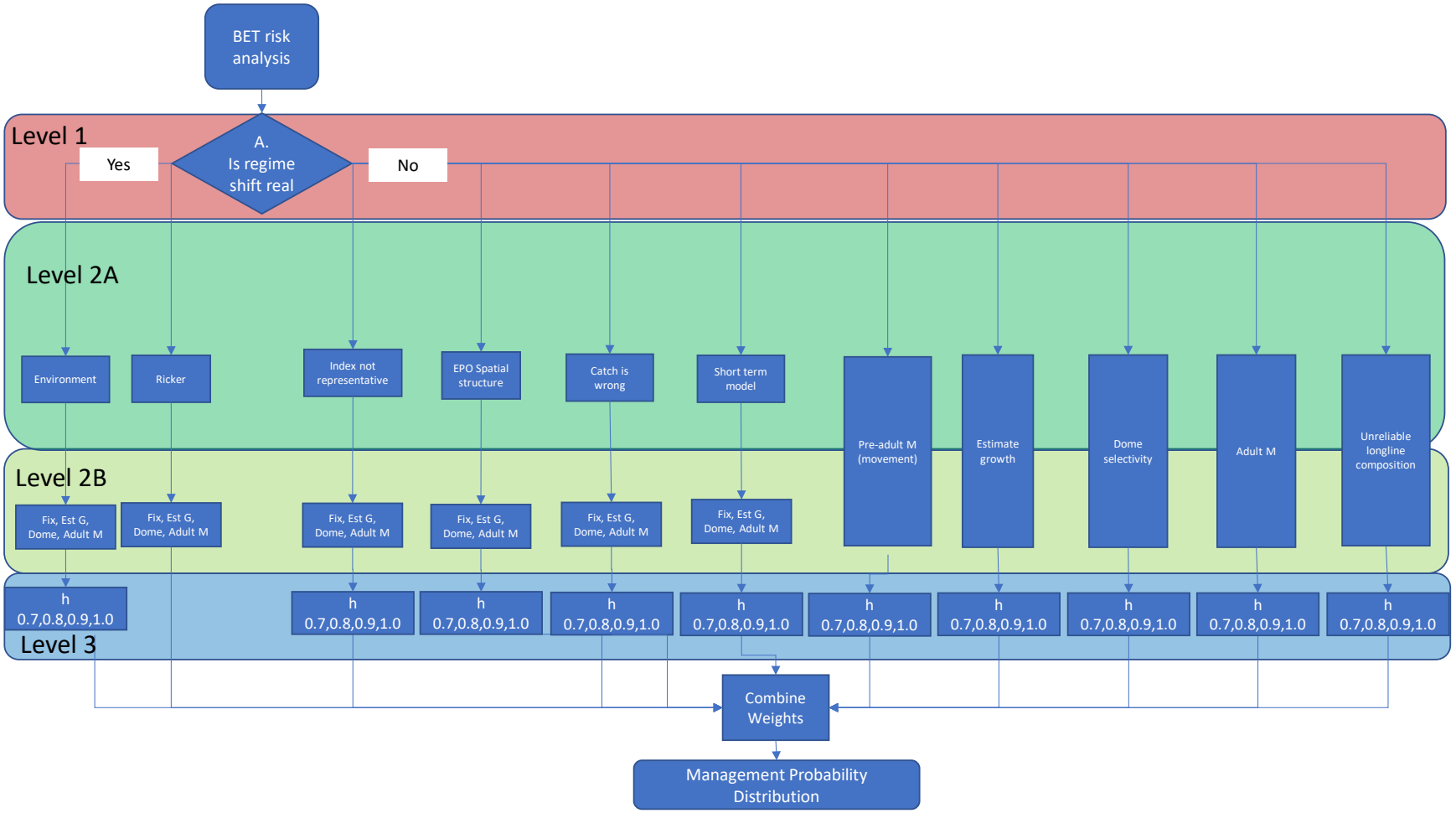


**FIGURE 2.** Flow chart describing the calculation of probabilities for the alternative hypotheses for bigeye tuna initially considered for the risk assessment.

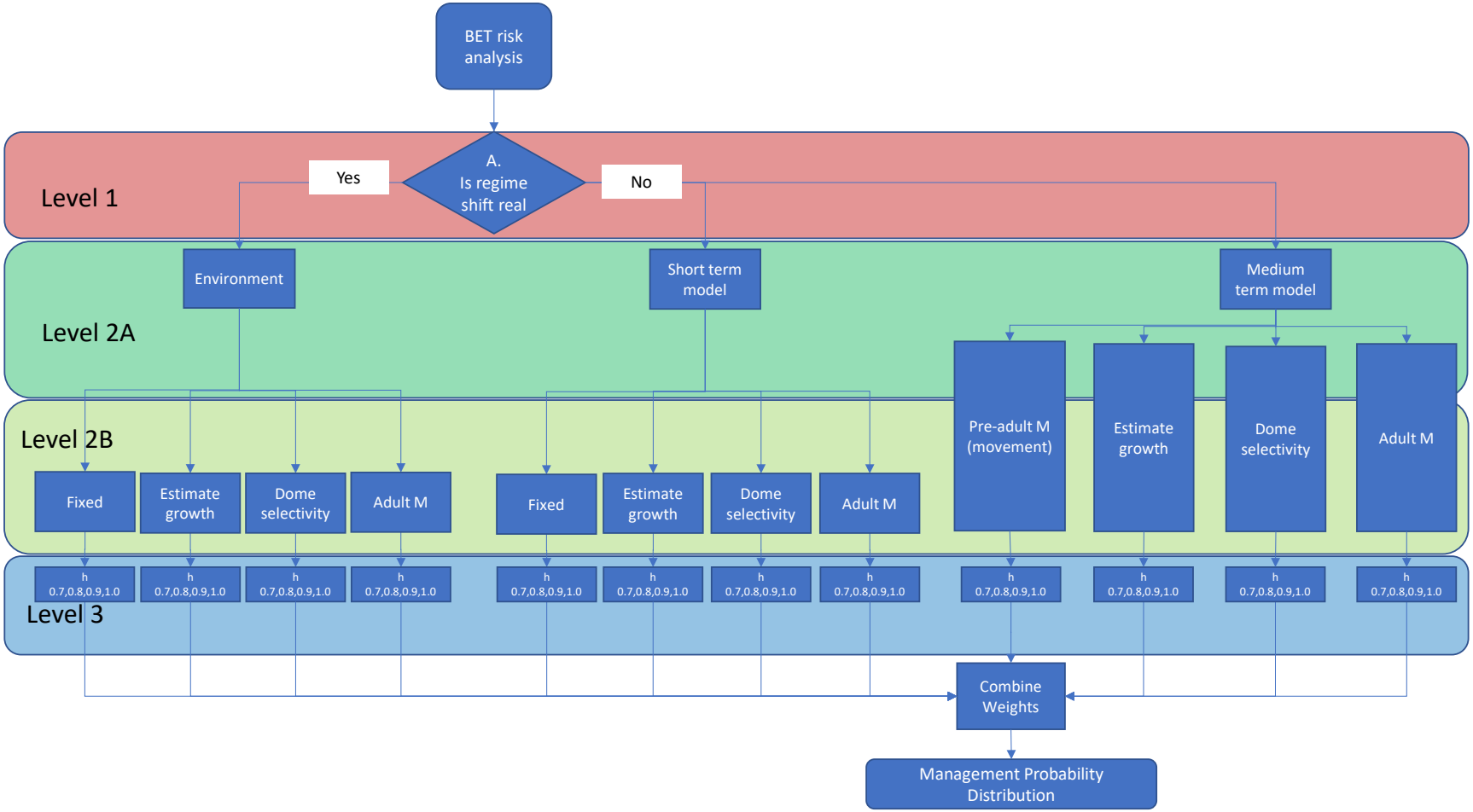


**FIGURE 3.** Flow chart describing the calculation of probabilities for the alternative hypotheses for bigeye tuna used in the final risk assessment.

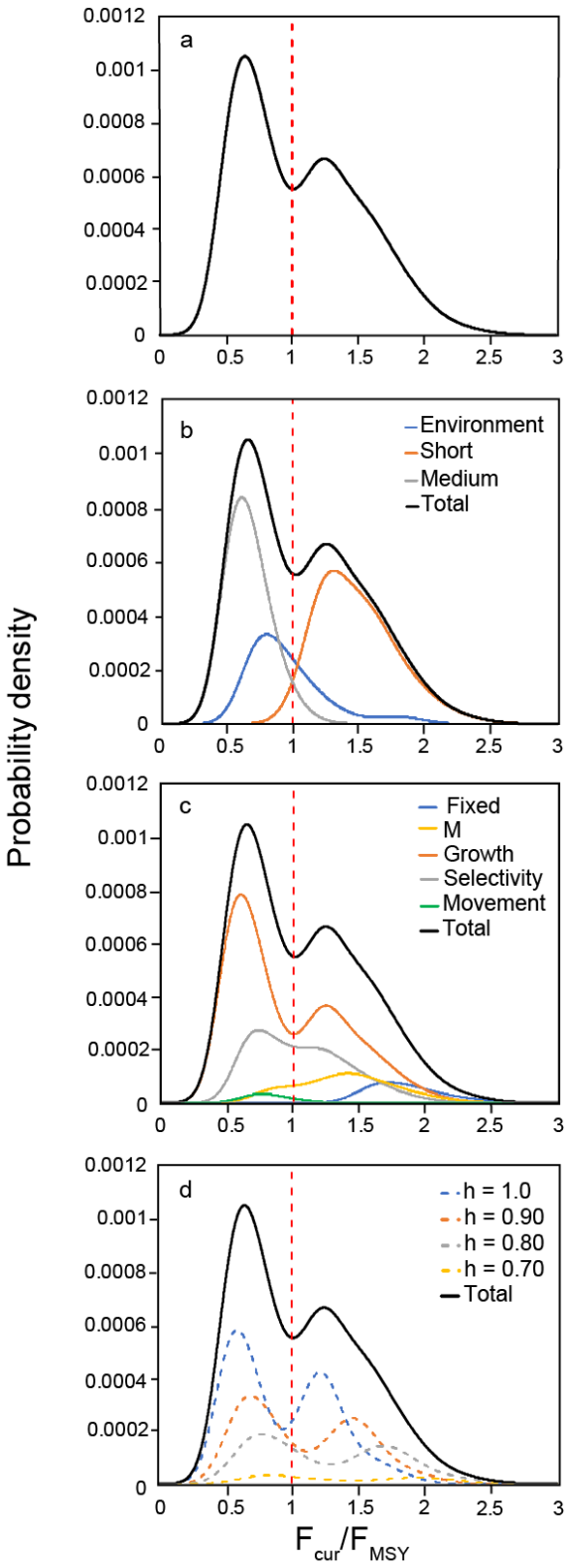


**FIGURE 4.** Probability density functions for $F_{cur}/F_{MSY}$ broken down into different components.

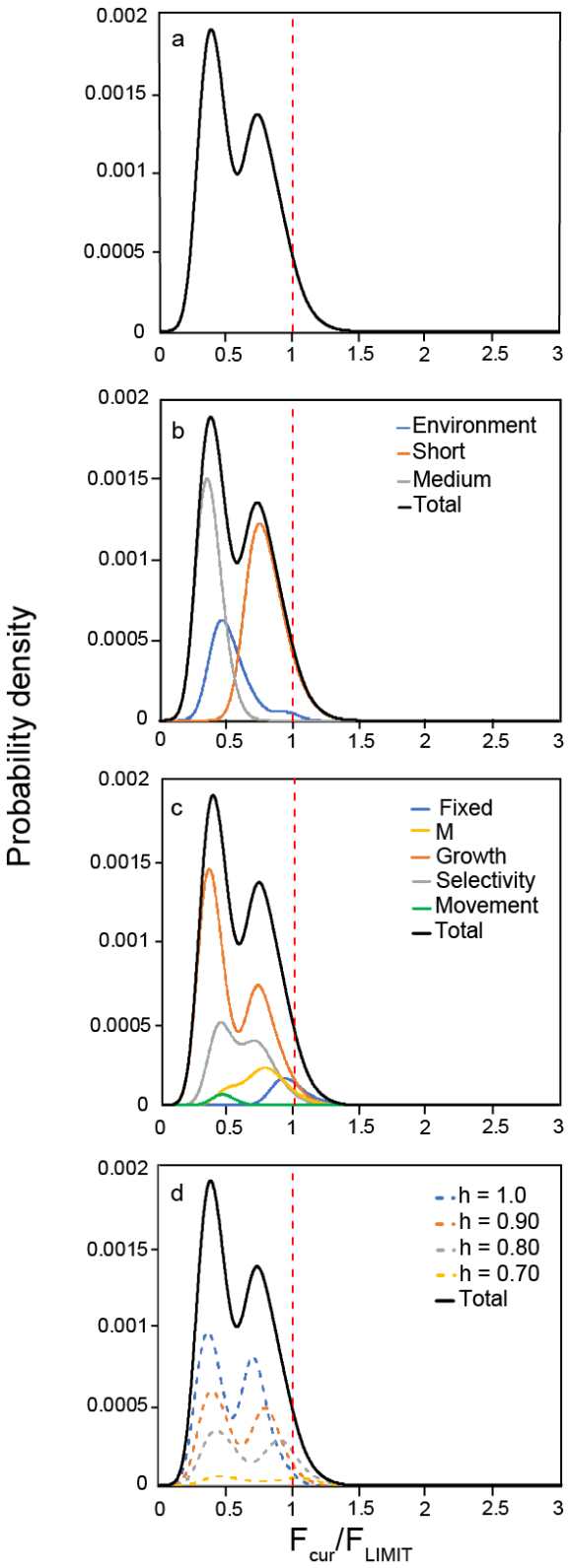


**FIGURE 5.** Probability density functions for $F_{cur}/F_{LIMIT}$ broken down into different components.

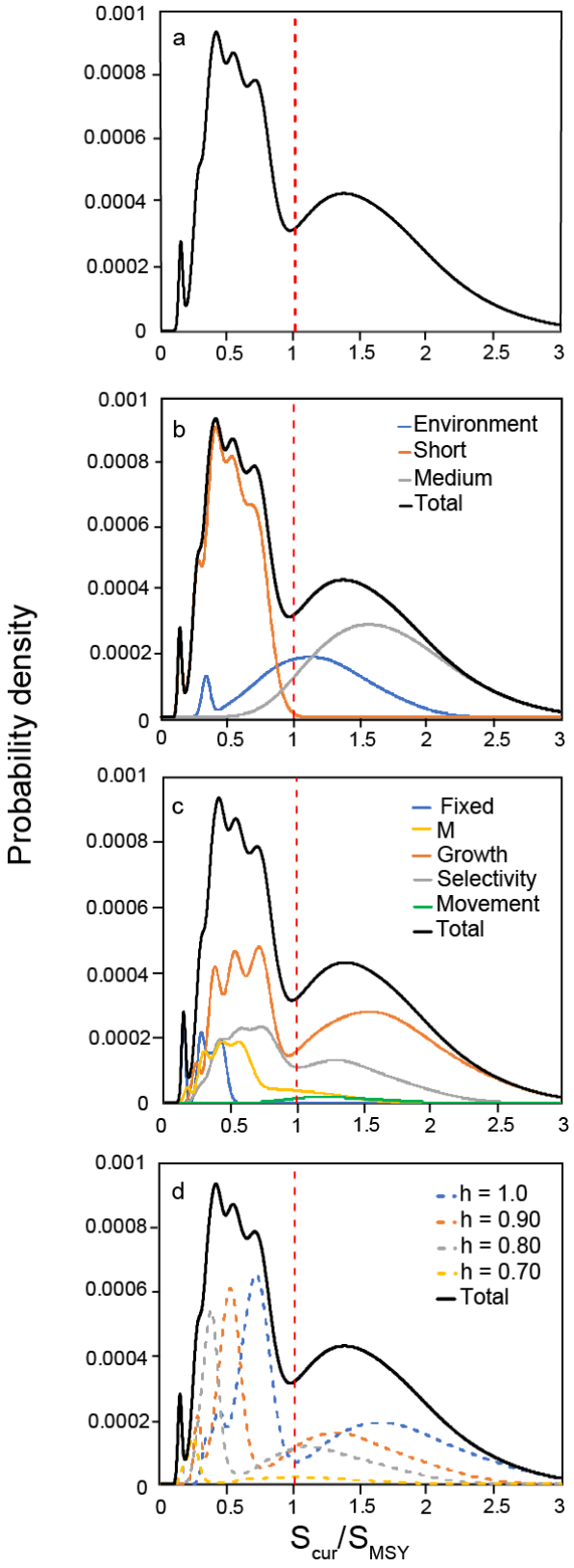


Figure 6. Probability density functions for $S_{cur}$ /$S_{MSY\_d}$ broken down into different components.

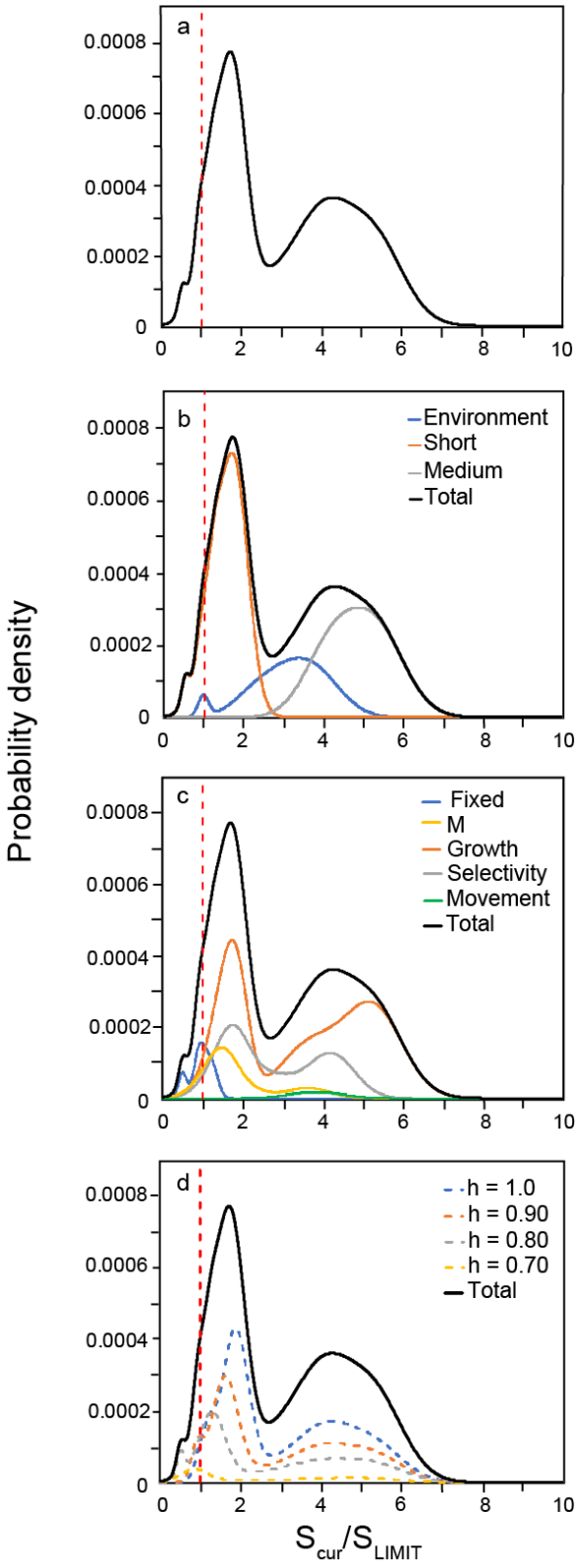


**FIGURE 7.** Probability density functions for $S_{cur}/S_{LIMIT}$ broken down into different components.

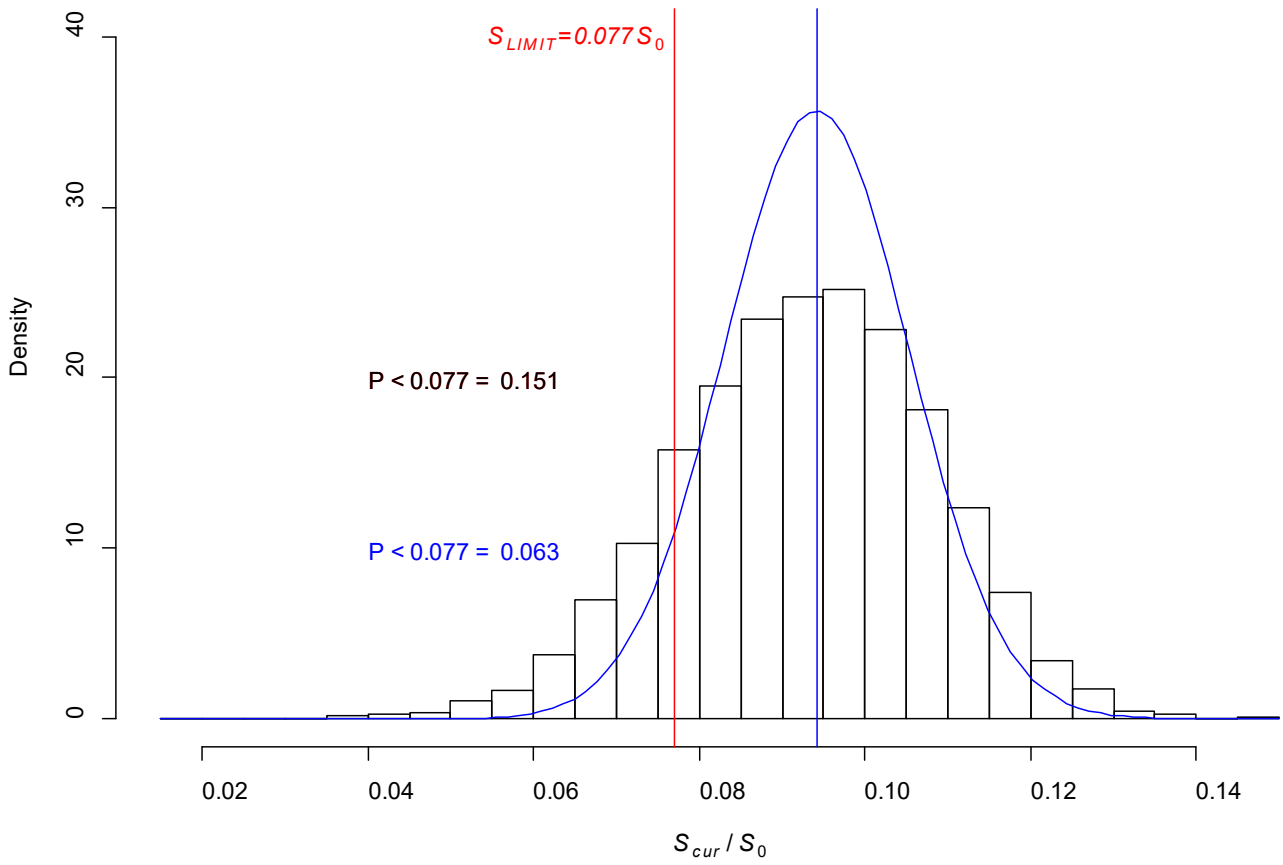


**FIGURE 8**. Bayesian posterior distribution (vertical bars) for the ratio of current biomass over equilibrium virgin biomass ($S_{cur}/S_0$) from MCMC of the bigeye *Short-term, Fixed* model. Red line is biomass limit reference point ($S_{LIMIT}$ =0.077$S_0$), blue vertical line is the mean $S_{cur}/S_0$ from the MLE, blue curve is the normal distribution from MLE estimates of mean and variance, "P < 0.077" are the probabilities of being below $S_{LIMIT}$ derived from both the MCMC posterior (black text) and the MLE normal distribution (blue text).