**INTER-AMERICAN TROPICAL TUNA COMMISSION**

# 2nd WORKSHOP ON IMPROVING THE RISK ANALYSIS FOR TROPICAL TUNAS IN THE EASTERN PACIFIC OCEAN: MODEL WEIGHTING IN INTEGRATED STOCK ASSESSMENTS

***by videoconference***

**28 November – 2 December 2022**

# CHAIR'S REPORT

Mark N. Maunder and Carolina Minte-Vera

**SUMMARY**

The Center for the Advancement of Population Assessment Methodology (CAPAM) and the Inter-American Tropical Tuna Commission (IATTC) held a virtual workshop on Model Weighting in Integrated Stock Assessments on 28 November to 1 December 2022 (the originally scheduled December 2 was not needed). The workshop focused on designing an approach to create an ensemble and to weight models in that ensemble that is more objective, transparent, and automated. This was part of the IATTC workplan to improve the risk analysis for tropical tunas in the Eastern Pacific Ocean (EPO), which was the approach used in the 2020 Benchmark assessments for yellowfin and bigeye tuna (SAC-11-06; SAC-11-Inf-F; SAC-11-07; SAC-11-Inf-J; IATTC-95-05). The workshop was conducted virtually (see here for recordings) for three hours each day over four days and generally followed the CAPAM workshop style, of presentations interspersed with ample discussion sessions. The workshop is relevant given the increasing use of ensembles in providing management advice and the lack of agreed good practices for weighting models. This report is structured around the following key questions that formed the basis for discussion: (a) Which models to consider and what measures (diagnostics) should be used to exclude models? (b) What measures to use in weighting and how to determine the weight for each metric? (c) How to combine weights? (d) How to present and use results?

Developing an ensemble is comprised of three steps a) what models should be considered? b) how to fix and/or eliminate models, and c) how to weight models. The models to consider should be based on the development of a conceptual model of the system and Good Practices to represent the alternatives identified by the conceptual model in the form of assessments models. Duplication of models should be avoided to prevent double weighting of certain hypotheses. Diagnostics are then used in an iterative process to fix and then accept or reject the models in the ensemble. Finally, the models in the ensemble should be weighted by either their ability to fit the data or to predict out of sample observations. However, the statistical properties of stock assessment models (e.g., inappropriate data weighting, unmodelled process variation, etc.) usually make fit to data an inappropriate measure for model weighting.

This report is not a consensus of the workshop participants, but an interpretation of the chair (Mark Maunder) and coauthors on the workshop presentations, discussions, and other available information.

## 1. INTRODUCTION

The workshop was part of the IATTC research program to improve the risk analysis for tropical tunas in the Eastern Pacific Ocean (EPO). Stock assessments are inherently uncertain, and this uncertainty should be taken into consideration when providing management advice. The IATTC, like many other management

agencies addresses the uncertainty through probability statements in its harvest control rule e.g. "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). At the same time, there is a movement away from single best assessments models to a set of reference models, an ensemble, that better represent our understanding of the system. In addition to providing management advice and evaluating probability statements, the ensemble can also form the basis for operating models used in Management Strategy Evaluation (MSE).

The IATTC risk analysis (SAC-11-Inf-F) uses an ensemble of models with each model representing an alternative hypothesis about the population dynamics and the data used to fit the models. Each of these models needs to be appropriately weighted based on its probability of being true. The model weighting in the IATTC risk analysis conducted in 2020 was mostly subjective and based on a range of metrics from model fit and diagnostics to convergence criteria and plausibility of results. The goal of the IATTC Stock Assessment Program is to design a more objective, transparent, and automated method for weighting fishery stock assessment model ensembles. This workshop series follows a request from IATTC SAC to improve the risk analysis for the tropical tuna (IATTC-97-01). We are using the CAPAM framework because it is a common challenge across all types of stock assessments and management. The first workshop on Diagnostics was held on January 31-Feburary 3, 2022 (WSRSK-01). This workshop series will continue at an ISSF sponsored session during the Tuna Stock Assessment Good Practices Workshop in New Zealand 7-10 March 2023.

The workshop was conducted virtually for three hours each day over four days (the originally scheduled fifth day was not needed) and generally followed the CAPAM style of workshop. It consisted of several invited speaker presentations on relevant topics and some contributed presentations. Ample time was provided after each presentation and in dedicated sessions at the end of each topic for questions and discussions. Chat facilities were enabled during the virtual meeting to encourage discussions and questions. The presentations and discussions were recorded and posted on the CAPAM website.

(https://www.iattc.org/en-US/Event/DetailMeeting/Meeting-WSRSK-02).

In preparation for the workshop, Nicholas Ducharme-Barth conducted a survey to determine how model weighting is being used in the fishery stock assessment community and the results were provided in his presentation. About 60% of the 42 respondents had used a model ensemble to characterize stock status. The Regional Fishery Management Organizations (RFMOs), including the Tuna RFMOs (tRFMOs) were well represented among respondents who had used ensembles. A variety of approaches have been used including a) ad hoc combination of models, b) hypothesis tree approaches, c) full factorial combination of uncertainties, and d) Monte Carlo Bootstrap (e.g., fixing parameters to values drawn from a pre-defined distribution). Nearly half of the ensembles combined distributions of the quantity of interest from each model to combine the ensemble and produce management advice.

The main deliverable of the workshop is this report. A majority of the information in the report comes from the presentations available here and recordings available here and the associated papers, which should be consulted for further information.

Most of this report summarizes the presentations of this, and other CAPAM workshops (see http://www.capamresearch.org/), and the published papers they are based on, and we encourage readers to use those resources and cite the corresponding papers rather than this report.

## 2. INTRODUCTION TO MODEL WEIGHTING

Carsten Dormann provided a review of model averaging and the way forward, focusing on ecological applications. The presentation was based on the recent publication "Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference" by Dormann et al. (2018). There are various reasons for using model averaging and the reasons may be different in ecology and fisheries than in other applications depending on the information content of the data available, the models used (e.g., predictive vs mechanistic), and the objectives of the analysis (e.g., the quantities being predicted). Often there is limited data, or the data is uninformative, so there is no best model. There may be uncertainty in initial or boundary conditions, or the model might include stochastic components, and averaging over this uncertainty may be desirable. Finally, there is empirical evidence that predictions from model averaging ensembles have lower long-term errors (i.e., mean squared error, MSE) than single models.

Dormann et al. (2018) show how separating the MSE into its components bias and variance can help understand how model averaging can improve predictions. The variance of an ensemble's predictions can be further separated into the variance of each model and the covariance among models. The weight given to each model is influencing these terms when calculating the model average. Therefore, three components are important for evaluating model averages: 1) bias, 2) variance, 3) covariance, and 4) model weights. Increased variance and bias increase the prediction error. Positive correlation among models increases the variance of the average and negative correlation decreases the variance of the average. The influence of weights will depend on the bias and variance of the models with increased weight compared to those that have decreased weight (the correlation will probably also have some impact).

It is generally considered that adding more models will reduce the total error, but this is not intuitive and not necessarily true. For example, using the correct model should produce the least error and adding additional incorrect models should only increase the error. However, given a model is always a simplification of the system, no model is correct. Also, adding perfectly correlated models (i.e., the same model) would not necessarily decrease the error. Adding unbiased models that are not perfectly correlated will reduce the variance of the mean prediction and if they are negatively correlated would produce a greater reduction in the variance. If the bias is randomly distributed around the truth, then adding biased models should reduce the variance of the mean prediction. However, we do not know the truth and therefore, particularly in process-based models, it is unclear if combining models will reduce the total error due to possible biases in the models.

There are numerous ways models can be weighted, but they can generally be grouped into a) Bayesian, b) information theoretic, and c) tactical. Information theoretic methods are generally applied to a few carefully considered models and therefore they have been considered to weight the models twice, once by expert opinion (i.e., which models to consider) and once by fit to the data. The tactical methods can vary widely and are often chosen because they make sense intuitively and/or they work (e.g., cross validation). Also, unlike the information theoretical methods, the tactical methods, which are typically used in machine learning, do not have to use likelihoods or adjust for the number of parameters.

The model weights could be estimated to produce the minimum error. However, this essentially adds "parameters" to the model and therefore increases the variance compared to fixing the weights at the correct values. The resulting error may even be larger than fixing the weights at approximately the right value or even equal weighting, which has also been shown for data weighting. It also should be noted that fixing the weights based on some procedure (e.g., expert opinion) ignores the additional variation due to data weighting, which may be relevant when consideration of uncertainty is a key component of management advice. Dormann et al. (2018) evaluated many different model weighting methods in a simple application (linear model) and found that most performed similarly including Bayesian model

averaging, equal weights, median, AIC weights, full model, and single best model. Those that performed noticeably worse tried to optimize the weights. In a more complicated application, they found that the methods tested often gave approximately equal weights to all models or selected just one model.

The final step is to combine the models together to represent the total uncertainty (e.g., estimate 95% confidence/probability intervals). There are a variety of ways of doing this, but the two promising methods are 1) using the full model (a model with all the covariates included and parameters estimated) and 2) combining the distributions of the quantity of interest from each model weighted by the model weight. In general, specifying and estimating the full model for a fishery stock assessment is not possible or computationally problematic. Therefore, combining distributions is probably the most appropriate approach. One issue with this approach is how to deal with models that have correlated predictions. The correlation can be obtained through cross validation, but this is complex to do for stock assessment models.

Dorman concluded that complex model weighting is not needed, and equal weights, median, or cross-validation are probably acceptable approaches as long as the models included are considered reasonable.

Michael Spence provided a more general view of model weighting. He explained how setting up model weights can be difficult because the models are often correlated since a) they are fitted to the same data, b) built with same knowledge, c) include similar processes, and d) fitted by similar or sometimes the same people. Model weighting can be viewed from the concept of discrepancy and how different models can provide different information. For example, if one model provides information on long term trends and another on short term fluctuations, or one on scale and the other on trends, then taking the appropriate average might improve predictions.

## 3. IATTC RISK ANALYSIS

Mark Maunder presented the IATTC Risk Analysis. Further details can be found in Maunder et al. (2020) (SAC-11-Inf-F). The hypotheses in the IATTC risk analysis were developed to address issues in the stock assessment rather than to represent alternative states of nature that were considered plausible. Although they did consider the main alternative hypotheses (natural mortality, growth, selectivity), a more thorough set of hypotheses based on a conceptual model should have been considered.

The IATTC Risk Analysis used a hierarchical approach to avoid model duplication and facilitate the construction and weighting of hypotheses. This also allowed for hypotheses representing parameters that cannot be estimated within the model by fitting to the data. For example, estimates of the steepness of the Beverton-Holt stock-recruitment relationship is inherently biased even if the model is correctly specified (Lee et al., 2012).

The first level in the Hypothesis Hierarchy includes Overarching Hypotheses that consist of broad states of nature (e.g., the number of stocks) and can be represented by a variety of models and data. These Overarching Hypotheses are not evaluated by fit to data and are given weights based on expert opinion. The second level of the Hypotheses Hierarchy address specific issues in the stock assessment each on its own sub-level. The sub levels are typically used in combination to solve all the assessment issues, but a specific hypothesis might solve more than one issue. The third level of the Hypothesis Hierarchy is evaluated differently to avoid the influence of the data, reduce the number of analyses, or for convenience. This level is typically encompassed by a single hypothesis which can be represented by restricting the model (e.g., by fixing the value of a parameter such as the steepness of the stock-recruitment relationship). The hypotheses on this level are applied to most, if not all, the hypotheses on level two.

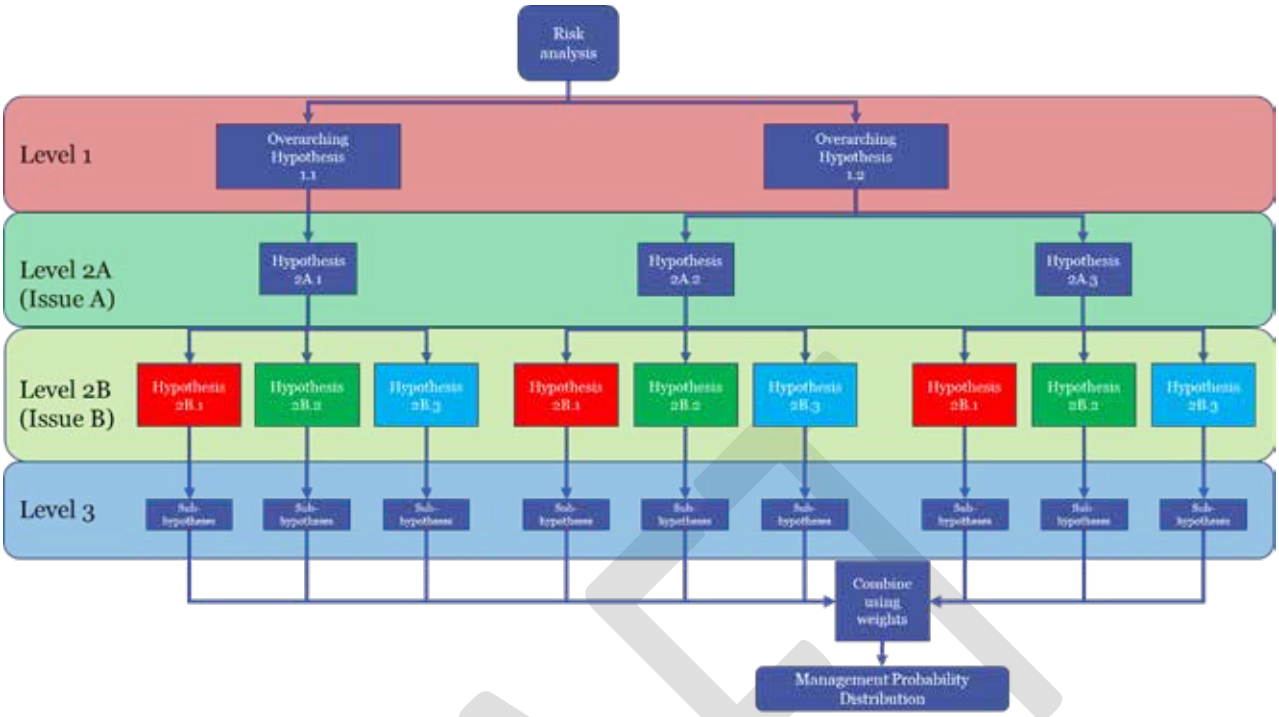


**FIGURE 3.** The Hypothesis Hierarchy used in the IATTC risk analysis.

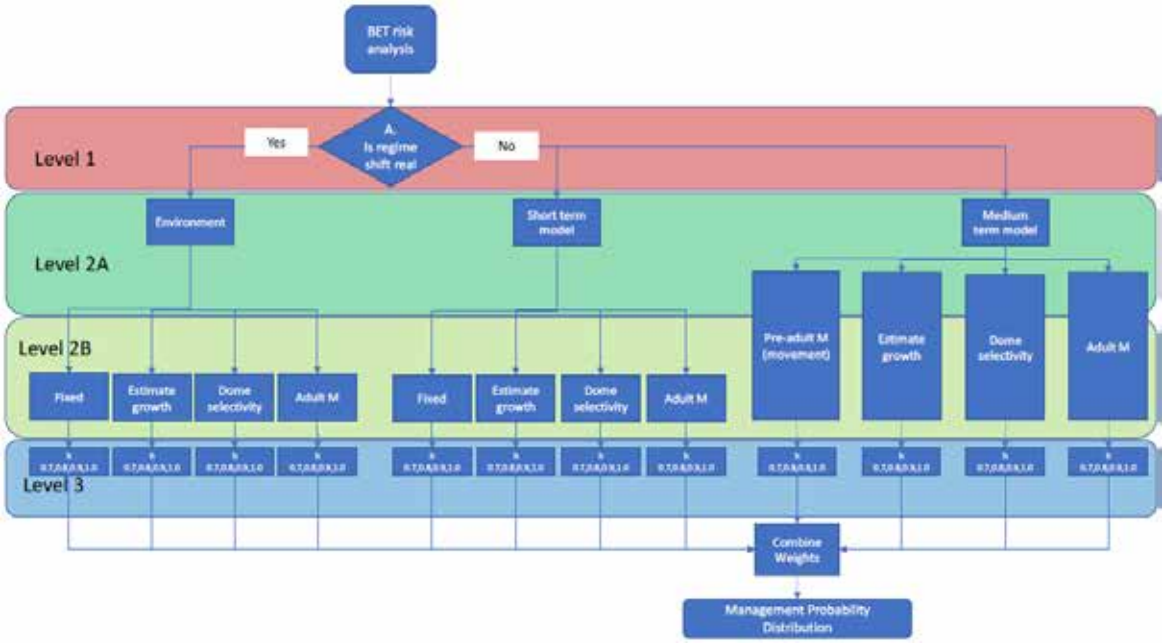


**FIGURE 4.** The Hypothesis Hierarchy used in the EPO bigeye Tuna risk analysis (need to get the power point to make the figure better).

Data weighting continues to be an issue with the stock assessments of tropical tunas in the EPO and therefore fit to the data is not an appropriate method to weight the models in the risk analysis. Several metrics were used to weight the models, with data fit only comprising a small portion of the weight. The data fit was further reduced by rescaling the range of the AIC. Other metrics used included expert opinion,

convergence, plausibility of parameter estimates, plausibility of results (e.g., estimates of fishing mortality and biomass), and diagnostics. The diagnostics used included the age-structured production model (ASPM), R0 likelihood component profile, catch-curve analysis, retrospective analysis, and residual analysis (composition, index, and recruitment). The risk analysis used a somewhat complicated approach to assign and rescale weights within a level of the hierarchy to avoid over weighting any single hypothesis.

The risk analysis was conducted by combining normal distributions of the quantity of interest from each model using the maximum likelihood estimate (MLE) and its standard error (SE). This produced a joint distribution that could then be used to evaluate the probability statements in the harvest control rule. The normal distribution approximation works well with informative data and can be checked by estimating the posterior distribution using MCMC. The results were presented as distributions, component distributions (distributions for groups of models represented by a level in the Hypothesis Hierarchy), cumulative distributions, and decision tables (IATTC-95-05).

The risk analysis was applied to yellowfin and bigeye tuna in the EPO (IATTC-95-05). The results for bigeye tuna (see Figure 4 for the Hypothesis Hierarchy) produced bimodal distributions with quite different results from different groups of models (Figure 5).

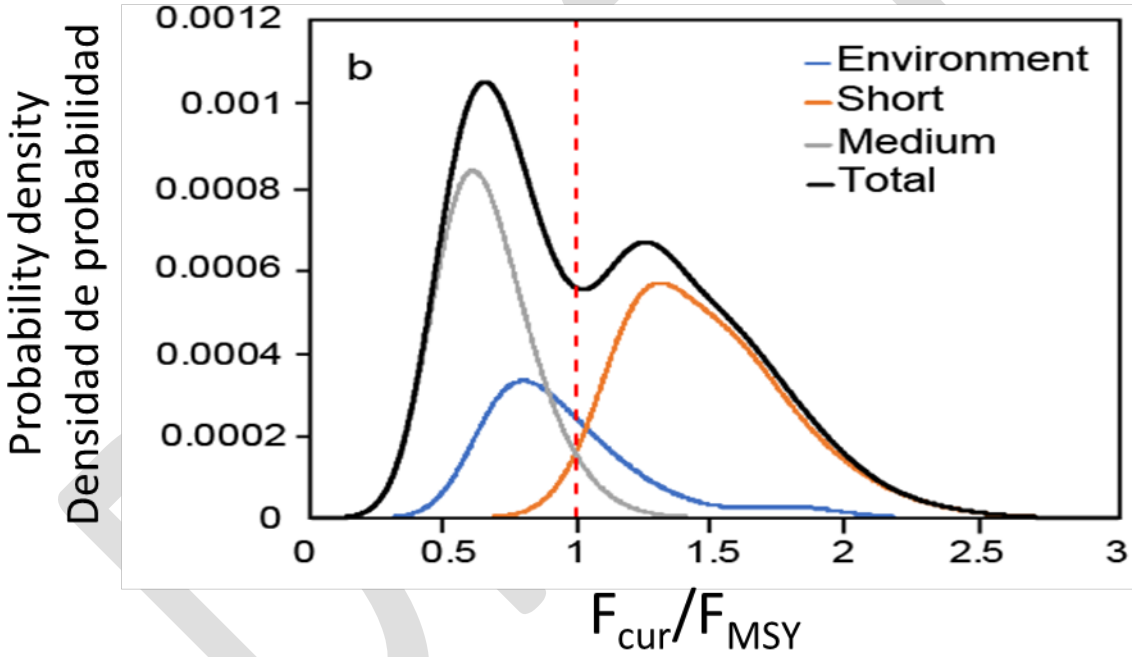


**FIGURE 5.** Estimated probability distribution for the fishing mortality relative to that corresponding to MSY $F_{cur}/F_{MSY}$ for the bigeye tuna application.

## 4. WHICH MODELS TO CONSIDER AND WHAT MEASURES (DIAGNOSTICS) SHOULD BE USED TO EXCLUDE MODELS?

The first step in building an ensemble is to determine which models to consider. This involves developing an understanding of the system and translating this into a stock assessment model. These models then need to be evaluated to determine which are appropriate representations of the system giving the data

and knowledge about the system. The first step generally involves developing a conceptual model of the system that encompasses understanding at all levels of the system including the population dynamics, the fishery dynamics, and the environment, as well as the data available (see the presentation by Carolina Minte-Vera). The conceptual model will often outline multiple hypotheses about different process that might represent the system. These hypotheses are then translated into a stock assessment model using Good Practices. The second step generally involves using diagnostics to evaluate if the model and results, including the fit to the data, do not violate the assumptions or if the assumptions are inappropriate.

### *4. 1.The conceptual model*

Creating a conceptual model is an important first step in creating an ensemble of models. A conceptual model (CM) is a simplified representation of the main processes and components of a systems and how they are related. It is composed of a collection of concepts intertwined in a logical way to represent the understanding and uncertainty about the functioning of a system. The CM is also a tool that can foster interdisciplinary collaboration providing a framework to integrated knowledge from diverse disciplines.

The conceptual model is used to identify the possible alternative hypotheses about how the system works. CMs are useful to organize ideas and rank the main processes and their uncertainties. The fishery system encompasses both natural (e.g., the target species, environmental influences, and ecosystem effects) and anthropogenic components (e.g., target and bycatch fishing fleets, scientific surveys, and fisheries management). It can be used to indicate which uncertainties are nearly impossible to reduce with data (*e.g.* steepness of the stock-recruitment curve for tuna stocks), can be reduced with the collection of alternative data sources (*e.g.* uncertainty on movement rates can be reduced by implementing good tagging programs). The CM provides a road map for the analysist to build the assessment model. It is also a useful communication tool among analysist, experts, stakeholders, providing a framework for knowledge integration from diverse disciplines.

The objectives of the assessment should be stated before building the conceptual model. The objectives will determine the scope of the conceptual model and will guide if there is enough information to proceed into operationalizing the conceptual model in a statistical assessment model. Some objectives include estimation of stock status, determination of total allowable catches, compliance with certification for sustainability which implies environmental/ecosystem impacts might be important, or to produce operating models for management strategy evaluation which might cover a larger range of processes).

The conceptual model should be based on knowledge. This knowledge includes looking at archetypes or general patterns from ecological theory: knowledge related to similar species in other ecosystems or other species in the same ecosystem, as well as general information from ecology theory, can be used to build the draft conceptual model. The available literature on the stock and species should be extensively reviewed. Some specific knowledge including using general patterns from fisheries science to inform the shape of the selectivity function, using exploratory data analysis to understand patterns in the data that may be related to the processes that are needed to be modelled. All data available can be used to construct the conceptual model regardless of whether it will be used in the assessment. Example of exploratory data analysis (EDA) include regression tree analysis on length frequency data, regression tree analysis on catch rates (CPUE), and spatio-temporal models for CPUE standardization.

The conceptual model building can also benefit from an elicitation process such a workshops with experts in related fields (*e.g.* biology, ecology, tagging) and stakeholders. Some components of a successful workshop include translating the current stock assessment assumptions into the implied conceptual model if an assessment already exists, an extensive literature review, performing the necessary exploratory data analyses (EDA), producing a strawman conceptual model based on the literature review

and EDA, identify the main experts to invite (biologists/ecologists working on the stock, biologists/ecologists working on the same species but other stocks, fisheries scientists, oceanographers, industry, and other stakeholder groups). The agenda and the discussion questions should be built around the strawman conceptual model.

The conceptual model can be represented in diverse ways as long as all the components are put together. The representations could include or be a combination of narratives, diagrams, maps, concept maps, mental maps. The components of the conceptual model should include responses to (or indication of uncertainty about) the following questions about:

1) spatial scale (latitude, longitude, depth)
    a. What is the distribution range of the stock?
    b. Is there uncertainty in the stock structure?
    c. Where are the spawning habitats / areas?
    d. Where are the nursery or juveniles habitats /areas?
    e. Where are the feeding habitats /areas?
    f. Is there movement of juveniles?
    g. Is there movement of adults?
    h. What are the underlying oceanographic forcing functions?
    i. Where are the fisheries operating?
    j. Are there heterogeneous conditions that may influence the catchability of the gear?
2) Temporal scale:
    a. What is the maximum age?
    b. What is the speed of growth?
    c. What is the age and or size of 50% maturation?
    d. Is seasonality important?
    e. When is data on catches available?
    f. When is data on indices of abundance available?
    g. When is data on length frequency available?
3) Production function:
    a. What is the growth pattern (mean and variation, CV of size at age)?
    b. Is growth sex-specific?
    c. What is the natural mortality pattern?
    d. Is natural mortality sex-specific?
    e. Does natural mortality vary by age or size?
    f. What is the recruitment function?
    g. What is the recruitment variability?

**Stock assessment cycle.** The traditional base case approach to stock assessment started with the analysist making assumptions that would translate into a draft model that would be subject to diagnostics (Figure 1). If problems were detected, the model would be modified to address them. That base-case model would be used to provide advice. Sensitivity analysis would be done to illustrate how changes on the main assumptions would affect the management advice. In the model ensemble approach, the assessment cycle starts with the construction of a conceptual model (Figure 2) which includes alternative hypotheses of how the system works. The analyst will then translate the conceptual model into assessment models. The first model to be built will be the "ancestral" model, which can be easily modified into models that represent alternative hypotheses. Areas where there is uncertainty will be translated into different models and the models will be subject to diagnostics and modified to address issues. The models that fail to present acceptable behavior on diagnostics would be discarded. The analyst should document each decision on how the hypotheses are translated into the models.

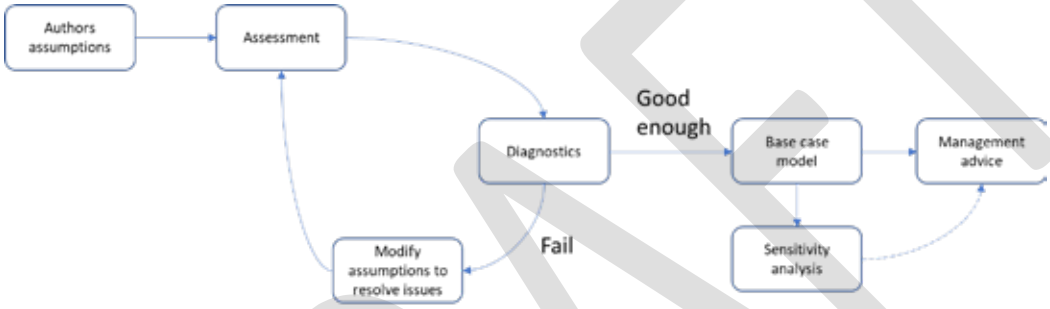


**FIGURE 1.** Historical base case approach to stock assessment

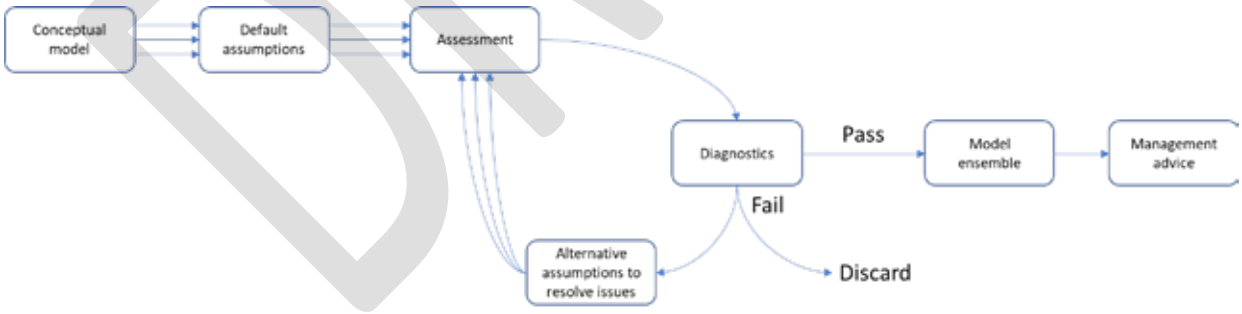


**FIGURE 2.** Model ensemble approach to stock assessment

### *4.2. Using diagnostics to fix and accept or reject models*

Mark Maunder provided a presentation on the use of diagnostics based on previous presentations from the CAPAM Stock Assessment Good Practices Workshop (http://www.capamresearch.org/GPG-Workshop) and the CAPAM Diagnostics Workshop (http://www.capamresearch.org/Diagnostics-Workshop). In general, a model that is used for management advice should represent the system being studied and not violate any of the assumptions about the system. Misspecifications could be 1) incorrect parameter values, 2) incorrect population dynamics model structure, 3) incorrect observation model structure, 4) ignoring process variation, 5) unrepresentative or poorly "standardized" data, or 6) incorrect specification of the likelihood functions. Violation of assumptions generally refers to the likelihood functions, which represent how the data is sampled, but since other specifications impact the model fit, they all can cause violation of assumptions. Model misspecification and violation of assumptions is typically evaluated using a variety of diagnostics tests. In fisheries stock assessment, both standard statistical diagnostics (e.g. residual analysis using runs tests) and test that have been specifically designed for stock assessment (e.g. the R0 likelihood component profile) are used. The IATTC/CAPAM workshop on diagnostics provides further details on each diagnostic (https://www.iattc.org/GetAttachment/30fc4743-0b40-4d73-b55b-2dda1d278980/WSRSK-01-RPT_1st-Workshop-on-improving-the-risk-analysis-for-the-tropical-tunas-in-the-EPO-model-diagnostics-for-integrated-stock-assessments.pdf). It concluded that "Model diagnostics provide tools to detect if there is a problem with the model, but current diagnostic tests can neither identify the exact source of the problem nor, if passed, guarantee that the model is an adequate representation of the "true" population dynamics."

A survey conducted at the diagnostic workshop found that goodness of fit (residual analysis), retrospective analysis, and the R0 component likelihood profile diagnostics were commonly used, but the other diagnostics including the age-structured production model, hindcasting, and catch-curve analysis were less frequently used. Although plausibility of results and parameter estimates were commonly used, they were probably subjective. The survey found similar support for whether these diagnostics should be used to evaluate the performance of the model and select models for use in management advice. The results were different for what diagnostics should be used for weighting models, with hindcasting and retrospective analysis the only ones with more than 50% support. The conclusions of the workshop report differed from the survey suggesting that all the diagnostics should be used to help diagnose model misspecification and select models to use in the ensemble. The workshop also concluded that "The development and understanding of diagnostics is not at the stage that diagnostics can be used for weighting models." But indicated that residual analysis, hindcasting and retrospective analysis had the potential to be used for model weighting. In general, the report concluded that current model diagnostics are good for model development, but less so for other purposes. They provide tools to detect if there is a problem with the model, but can't identify the exact source of the problem and do not guarantee that the model is an adequate representation of the "true" population dynamics nor whether the estimates of management quantities are reliable.

In general, the criteria to determine if a diagnostic has failed are limited and often rely on visual and subjective evaluation. There are some quantitative metrics, but many of these do not have criteria to determine failure or if they do, they are very subjective. The criteria for many of the diagnostics are likely to be application specific. Data that could be used to determine some criteria, such as plausibility of parameter estimates, should be used in the model or to create priors. Some diagnostics require a particular specification to detect certain model misspecification. For example runs tests of composition data could be applied over a) age/length or consecutive groups of ages/lengths to identify misspecified selectivity curves, growth model, or other process, b) over a year or block of years to detect unmodelled changes in selectivity, growth, or other processes, or c) over cohort to identify cohort targeting or cohort-

specific growth or other processes. Patterns in residuals may indicate unmodelled temporal variation in system or sampling processes. However, allowing variation in one process can eliminate residual patterns caused by time-variation in other parameters. A SDNR > 1 might indicate that the input sample sizes have not correctly accounted for the way the data were collected or the model is too stiff, while SDNR < 1 might indicate that the sample size was based on the wrong measure (e.g. tows sampled). Some diagnostics are intuitively appealing, but their performance has not lived up to expectations. For example, the R0 likelihood component profile visually shows data conflict, which indicates model misspecification, but application of Wang and Maunder's (20xx) metric has been shown to have low power to detect model misspecification (Carvalho et al. 2017) and the conflict may occur in data not directly associated with the misspecification. Other diagnostics, such as the age-structured production model, may indicate the information content of data, but not necessarily that the model is misspecified. Similarly, the catch-curve analysis indicates problems when none exist (Carvalho et al. 2017), but intuitively detects changes in selectivity and conflict between indices of abundance and composition data.

Some diagnostics have been studied more intensively and have specific metrics and failure criteria that are used. For example, ICES uses the range [-0.15-0.2] as indicating an acceptable model for Mohn's Rho of a retrospective analysis (ICES, 2019). Values outside this range indicate that there might be errors in the catch time series or processes are time varying but not modelled. Some advice on interpreting retrospective analysis are that a single large error can be ignored, large but random errors indicate that the estimates are uncertain and the uncertainty should be taken into consideration when providing management advice, moderate to large pattern requires the model to be fixed. It should be noted that adding time varying process may reduce retrospective error, but may not improve the management related quantity (Szuwalski et al. 2018).

**TABLE X.** List of diagnostics, their metrics and failure criteria. Note that most of the criteria provided are very preliminary and subjective, they are likely to be application specific, and further research is needed to specify them.

| Diagnostic type | Metric | Criteria |
|---|---|---|
| Convergence | Hessian is positive define | NA |
| | Gradient | > 0.1 |
| | Parameter close to bound | < 1% |
| | Parameter CV | > 1.0 |
| | Parameter correlation | > 0.5 |
| Plausible results | Spawning biomass depletion | < 5% |
| | F | > 2, < 0.05 |
| Plausible parameter value | Various (e.g. M, K, Linf, h, etc.) | |
| Residual analysis | Runs test | |
| | SDNR | > X, < Y |
| Likelihood component profile | Wang and Maunder's (20xx) metric | |
| Retrospective analysis | Mohn's Rho | > -0.15, < 0.2 |
| | | |
| | | |

There are two main approaches that could be used to fix and accept or reject models using diagnostics. The first approach takes each model representing hypotheses from the conceptual model and applies the

diagnostics. If the model passes all the diagnostics, it is included in the ensemble. However, if the model fails the diagnostics, it is modified based on the outcome of the diagnostics with the aim to fix the model so it passes all the diagnostic and can be included in the ensemble. Multiple models could be developed and accepted for each hypothesis. The process is repeated until the diagnostics are passed or the model cannot be fixed and is rejected.

The second approach is to develop one (or a few) reference model that is thought to best represent the system and apply the diagnostics to that to create a good model. Models representing alternative hypotheses are then derived from the reference model. Many different alternatives in combination are developed and the diagnostics are applied to these models, eliminating models that do not pass the diagnostics. In this approach, rather than evaluating each individual model and trying to fix it, it is hoped that the combination of a large number of alternative hypotheses will create acceptable models that pass the diagnostics.

## 5. WHAT MEASURES TO USE IN WEIGHTING AND HOW TO DETERMINE THE WEIGHT FOR EACH METRIC?

Philipp Neubauer reviewed the different approaches to weight stock assessment models and illustrated them with an application to SW-Pacific blue shark. These approaches vary from giving all the weight to a single base-case model to giving equal weight to numerous models, often based on a grid of parameter values in full combination. In theory, all the components of model weighting could be integrated into the stock assessment, but in practice this is not possible. The appropriate approach is likely some intermediate between the two extremes that is practical to implement. Base-case assessments are usually chosen by expert opinion, diagnostics, or a combination of the two, and usually in an informal ad hoc way. The grid of parameters is usually also decided by expert opinion and/or initial model evaluation using diagnostics. The intermediate procedure should apply model weights and take into consideration a) alternative datasets/sources, b) alternative model structures, c) prior formulations, d) model diagnostics, e) model fit, and f) model adequacy for management advice.

The IATTC risk analysis method, as presented by Mark Maunder and described in Maunder et al. (2020), uses a variety of metrics to weight the models, including diagnostics, because the statistical properties of stock assessment models (e.g. inappropriate data weighting, unmodelled process variation, etc.) usually make fit to data an inappropriate measure for model weighting. However, the diagnostic workshop concluded that "The development and understanding of diagnostics is not at the stage that diagnostics can be used for weighting models. This is partly because current metrics from the diagnostics (e.g., Mohn's rho from retrospective analysis and MASE from hindcasting) cannot be turned into P(Model) or made consistent with AIC.". Therefore, there are two obvious ways forward, 1) improve the statistical properties of stock assessments so that measure of fit can be used as model weights or 2) improve out of sample prediction methods (e.g. cross validation or hindcasting) so they can be used for model weighting.

### *5.1 Model fit*

There are several properties of stock assessment models that make fit to data an inappropriate measure for model weighting. These may include a) multiple model assumptions are possible and may use different data; b) they are complex and highly parameterized; c) models are misspecified, d) penalized likelihood approaches are used to implement random effects, e) inappropriate data weighting, and f) unmodelled process variation. Some of these properties can be improved following the Good Practices identified at the CAPAM workshop on Stock Assessment Good practices. These include changes such as 1) appropriately implementing random effects through integration, 2) including process variation on more processes (e.g. fishery selectivity), and 3) developing data weighting by analyzing the sampling process.

However, it is not clear if the statistical properties of the assessment model can be improved such that data fit (e.g. AIC) can be used for model weighting.

### *5. 2 Diagnostics*

Diagnostics can be used for a variety of objectives (e.g. model selection, model weighting, characterizing uncertainty, data selection, determining value of information, and determining stakeholder information) but are typically used to detect if the model assumptions have been violated and are used to reject a model. However, it is not clear which diagnostics and how bad they have to perform before the model is not useful for management advice. For most diagnostics there is no clear metric or rejection criteria. Therefore, using diagnostics as weights might be a reasonable approach as a quasi-rejection method such that a model is essentially eliminated by giving the model zero weight when the diagnostic is bad, but not eliminating the model completely when the diagnostic is not so bad.

### *5. 3 Cross validation and hindcasting*

Cross validation measures the ability of a model to predict data that have not been used to estimate the model's parameters. The better the model is at predicting out of sample, the better it should be at predicting new data. In general, but not necessarily, this means that the model is a better representation of the system and therefore should provide better estimates of management quantities. In some cases, this may be a large leap of faith, but since we typically do not observe the management quantities, this is an assumption we must make.

Population dynamics models are time series and therefore make applying cross validation more complicated. In general, simply randomly selecting data for the training data set and the test data set is inappropriate and one-step-head prediction is necessary. Hindcasting, a version of one-step-ahead prediction, is becoming popular in fisheries stock assessment. There are several decisions that have to be made including the data type and number of years to predict, what data and how many years to include in the assessment, and how to measure the prediction. We refer you to Kell et al. (2016; 2021) for details about much of the information summarized in the presentation by Mark Maunder and also given below.

There are a variety of observed data types that could be predicted, but it makes sense to predict the data that is closest to the management quantity of interest. In general, this is probably an index of abundance that represents the spawning biomass. However, other measures such as the mean length of that index may also be informative. Mean length might be a better measure than the length composition because it accounts for correlation in the composition data (c.f., Francis 2011) and it is easier to define a metric. In some cases where prediction of a recruiting cohort is important for the management quantity of interest, predicting the mean length of a fishery that catches small fish might be useful. The data to predict could also be subdivided or grouped by several characteristics such as series, data type, fleet, time blocks, individual points, or a combination of these. These breakdowns are probably more appropriate when using hindcasting as a diagnostic to evaluate data conflicts. More research is needed to determine the right quantity to predict for each management quantity of interest.

When doing a peal (removing one time increment (e.g., year) of data) there are several decision to make. First is how many years each peal represents. A single time period is probably most appropriate, but it might also be related to the management cycle. Another decision is what data to leave in the model. Obviously, the data to be predicted needs to be left out of the model, but there are usually other data types that are used in the model, and should this data be left in the model for the years that are predicted? The final decision about the data is how many years should be predicted, should it be all years to the end of the models time frame or should it just be one step ahead.

The final decision is what metric should be used. There have been several different metrics used or proposed including root mean squared error (RMSE), which is sensitive to outliers, correlation, Mohn's rho (used for retrospective analysis), relative error, and mean absolute scaled error (MASE, see Kell et al., 2021). MASE compares the prediction to a naive prediction (last years value) and has the characteristic of scale invariance, symmetry, interpretability, and asymptotic normality. However, converting the MASE to an appropriate value for model weighting has been questioned. The obvious alternative is using the likelihood. However, this would require determining the appropriate data weighting (i.e. estimating the variance parameter).

## 6. HOW TO COMBINE WEIGHTS.

Allan Hicks presented considerations when combining models in an assemble and provided examples from Pacific halibut. He also explained how ensembles are used for the operating model (OM) in management strategy evaluation (MSE) and contrasted using an ensemble for management advice versus using MSE. Pacific halibut has a relatively long history of using ensembles compared to other stocks (since 2012). Only a limited number of models are used (4) and they are weighted equally. The distributions for management quantities are created by combining samples from the multivariate normal distribution for each model and combining them. The approach can be applied with or without using the covariance. This is like the approach used in the IATTC risk analysis, except the IATTC uses a grid approach to calculate the normal distributions and does not account for the autocorrelation among management parameters.

Hicks emphasize the fact that creating a combined distribution of the underlying quantity (e.g. F) and then comparing it to its reference point (e.g. $F_{MSY}$) is not the same as creating a combined distribution of the underlying quantity relative to its reference point (e.g. $F/F_{MSY}$). The IHC has created a R package to implement ensemble models that is available to other assessment authors.

Nicholas Ducharme-Barth presented the key decisions required when constructing and combining ensembles. Much of the presentation came from the paper "Focusing on the front end: A framework for incorporating uncertainty in biological parameters in model ensembles of integrated stock assessments" by Ducharme-Barth and Vincent (2022). He illustrated the key decisions using applications to SW Pacific Ocean swordfish and N Pacific Ocean blue shark. Full factorial approaches are inefficient, which can result in many models with unrealistic parameter combinations, and the choices for parameter levels and model weighting are often subjective. Although, the full factorial approach may be useful in applications with few discrete choices, a better approach might be to develop a multivariate distribution for key parameters, particularly continuous parameters, that would be sampled from and fixed in the assessment. This Monte Carlo Bootstrap (MCB) approach preserves parameter correlation and uncertainty and provides implicit weighting from the distributional assumptions. The multivariate distributions can be informed from life history relationships or other information. Censoring unlikely combinations can further reduce uncertainty. For the swordfish example, Ducharme-Barth found the results from the MCB approach differed from the full factorial approach in terms of both central tendency and uncertainty and required much fewer models to be run to characterize reference points. He also noted that care needs to be taken when the tails of a probability distribution are used in management advice, which is commonly the case, because they are often more dependent of how the models are weighted.

In the SWPO swordfish application, including estimation uncertainty produced some increase in uncertainty, but did not meaningfully change risk relative to reference points. Weighting the models by the fit (likelihood based) did not change the distributions. In the NPO blue shark application, there was a noticeable increase in risk when estimation uncertainty was included.

The hypothesis hierarchy approach was also discussed during the workshop. This approach can be directly linked to the conceptual model and the hierarchy can be tailored to remove redundant and/or unlikely combinations. The MCB, full factorial, and hypothesis hierarchy approaches can be combined applying each approach to different components of the model as appropriate.

Matthew Vincent explained how Monte-Carlo Bootstrap (MCB) could be used to produce model ensembles and illustrated it with application to SE Atlantic stock assessments (Restrepo et al., 1992; Legault et al., 2002, SEDAR 2022). Representing uncertainty is important because the acceptable biological catch (ABC) is based on reducing the overfishing limit (OFL) determined by the model by an amount associated with scientific uncertainty using the P* Approach. Many of the SE Atlantic stocks have large uncertainty in recreational and/or commercial historical landings and the recreational harvest includes discards of fish where the fate of these fish is not well understood for most species. The MCB can be used to sample from the uncertainty about the landings so that it is included in the total uncertainty of the model ensemble. Few species have estimates of natural mortality or sufficient age composition samples to estimate M and estimation of steepness is difficult due to flat likelihood profiles or inherent bias. Both M and steepness can be highly influential on estimates of management quantities. The MCB samples both data sources and parameters from their distributions. The models are not evaluated by diagnostics, but those that do not converge, with parameters estimated close to bounds, and some with extreme values of biomass or $F_{MSY}$, are removed. The models are given equal weight, but there is implicit weighting of data and parameters based on the distributions used in the MCB. Projections are conducted with stochastic recruitment. The past application of this approach has not accounted for uncertainty in estimates of stock status from each model in the ensemble (i.e. only model uncertainty), it has not incorporated models with structural differences, and typically assumes independent distributions for each fixed parameter, but these could be combined the overarching framework of other ensemble configurations (e.g., a hierarchical approach).

Max Cardinale illustrated model weighting of ensemble models using three case studies: Northern shrimp, Adriatic sole, and Gulf of Bothnian vendace. He focused on using diagnostics to weight the models. The Adriatic sole example used a hierarchical approach like the IATTC risk analysis with differences including data sets used, growth curves, and selectivity assumptions. Some models were eliminated because of poor diagnostics and others because the results were like other models. The estimated probability distributions of the quantities of interest for each model were combined (stitched) together using the model weight. The weighting was based on convergence (positive definite Hessian and jittering), goodness of fit (joint residuals and runs tests), consistency (retrospective analysis), and prediction skill (hindcasting). The final weights were essentially equal weighting for all models.

The Bothnian Sea vendace used a full combination of three factors, each having three levels: a) seal predation, b) basal natural mortality, and c) steepness of the stock-recruitment relationship. The final weighting was based on the same criteria used for Adriatic sole. The weighted median of plausible scenarios was used rather than stitching the distributions together. All except two of the models had the same weight.

The northern shrimp models used the hierarchical approach with the models derived from an “ancestral” model. The hypotheses included sex ratio, length of landings time series, hermaphroditism, and natural mortality. Some models were initially discarded because they had poor fit. A substantial number of models had poor diagnostics and were excluded from the ensemble (given zero weight). The weighted median of plausible scenarios was used rather than stitching the distributions together.

Cardinale made several conclusions about model weighting. First, the development of a reference model is the key aspect of the ensemble process as it can be used to derive the other models in the ensemble.

The weighting scheme must be agreed beforehand to avoid cherry picking of models by the participants.

A pass/fail system works well for stock assessment models for which diagnostics are used to weigh the models in the ensemble because differences in diagnostics performances between models is often small. Combining (stitching) the distributions of management quantities of each model is important to better represent the tails of the distribution and quantify risk. Finally, simulation testing is needed to determine which diagnostics should be used as weighting metrics.

## 7. HOW TO PRESENT AND USE RESULTS

Presenting results is an important part of using model weighting for management advice. There are a variety of ways the results can be presented from simple decision tables to graphical representation of the complete distribution and its components. However, the results have to be presented in a way that the users (e.g. managers) can understand. This often means hiding some of the important details. For example, a simple decision table might hide the bimodal distributions seen in the EPO bigeye tuna example. Currently there is no clear advice on how to present the results, but plotting the full distribution with components, possibly grouping the components into logical groups, should probably be included with any advice to ensure the users understand that multiple outcomes with different consequences are possible.

## 8. A PROPOSED APPROACH FOR MODEL WEIGHTING FOR TROPICAL TUNAS IN THE EPO

Carolina Minte-Vera presented the outline of an approach for model weighting for tropical tunas in the EPO. This approach was based on the original risk analysis applied to bigeye and yellowfin tuna in the EPO (SAC-11-06; SAC-11-Inf-F; SAC-11-07; SAC-11-Inf-J; IATTC-95-05) modified based on subsequent research and information presented at this workshop. The approach starts with developing a conceptual model of the system. The conceptual model is a broad description of the system that can then be used to develop the stock assessment models that represent the system and the alternative hypotheses, while being practical within computational constraints and being able to address the management questions. Development of the conceptual model relies on various tasks ranging from literature review to exploratory analyses (e.g., regression tree analysis of composition data to identify fisheries) and covers a wide range of topics from population dynamics and ecological interactions to oceanography and fisheries characteristics. Some factors that need to be considered because they are important for defining the stock assessment model include spatial and temporal scales (e.g., time step, maximum age, starting year) of the population (e.g., stock structure) and fishery (e.g., fishery definitions) dynamics, and population biology (e.g., natural mortality, growth, reproduction, stock-recruitment relationship, movement). An elicitation process should be conducted including workshops and consultation with experts and various stakeholders. The conceptual model should be checked for logical consistency, which might involve an independent review.

Alternative hypotheses should be constructed about components of the system for which there is incomplete knowledge and justification provided for each. Hypotheses should be identified as those that are practical and those that are impractical (i.e., can't be implemented given the available data/tools).

Impractical hypotheses will have zero weight, and could be considered in the future when the necessary resources become available. The hypotheses should be organized in a hierarchical system providing independent (orthogonal) uncertainty axes. Processes with no information in the data used to fit the model about their parameters or inherent bias should be identified and represented with alternative fixed

values (Level 3 hypotheses in figure 3). These can be identified from published simulation studies (e.g., Lee et al (20, 2012) for M and h) and/or specific simulations for the system under study. Processes with parameters that can be informed by external data should use that information, which could be included in a variety of ways i) fixed value, ii) parameter range, iii) prior distribution, iv) joint prior distribution for correlated processes (e.g., natural mortality and growth), v) the data integrated directly into the model if possible, or vi) as a range of fixed values with the information used in the weighting system.

There are several approaches to set up the stock assessment models that represent the alternative hypotheses. But in general, it is best to start by creating a reference ("ancestral") model based on the most likely assumptions after applying extensive diagnostic analysis to evaluate the model. The reference model and/or conceptual model can be changed when diagnostics highlight issues that merit new hypotheses. Models representing the alternative hypotheses will be developed from this reference model. The models can use different data or have completely different structure according to the hypothesis they represent.

Two approaches to evaluate the alternative hypotheses are considered. The first approach evaluates many models with alternative assumptions (e.g., different values of model parameters such as the steepness of the stock-recruitment relationship) or different model structures (e.g., alternative forms of the growth curve) in combination and then applies diagnostics to each model and if any diagnostic fails the model is rejected. This approach assumes that the combinations of model assumptions are enough to fix any model misspecification. The second approach starts with the best model for each hypothesis and then iteratively applies the diagnostics with modification of the model when it fails a diagnostic until the model passes all diagnostics, or there are no more appropriate modifications and the hypothesis is rejected. Multiple models could be created to represent each hypothesis or to correct a poor diagnostic. The diagnostics and their rejection criteria need to be determined in advance, noting that some diagnostics are for model understanding rather than for model selection. Data will need to have their own hypotheses, diagnostics, and quality control rules to be able to be included in the model.

Diagnostics that need to be passed for a model to be included are 1) model convergence (e.g., size of the gradient at the MLE, Hessian matrix is positive definite, jitter analysis), 2) no parameters on the bounds, 3) residual runs tests taking into account the assumed distribution and using PIT residuals for composition data, 4) retrospective analysis, 5) R0 likelihood component profile, 6) catch-curve analysis, 7) empirical selectivity, 8) plausible parameter estimates, and 9) plausible results (e.g., small F for highly exploited stocks). The metrics and criteria for pass/fail for each diagnostic still need to be worked out for most, if not all, diagnostics.

The weighting scheme should be decided in advance based on the objectives of the assessment. Expert judgement will be used to weight overarching hypotheses that cannot be evaluated by the application of models (e.g., stock structure). Similarly, for processes not informed by the data or inherently biased (e.g., steepness, movement rates), prior distributions (e.g., from meta-analyses, expert opinion or other) are used as weights. The remaining hypotheses are weighted by predictive ability (hindcast). Ideally the predictive ability should be estimated on the quantity of interest. However, quantities of interest in stock assessment are in general not directly observable. Therefore, the observable quantities should be chosen carefully (prefer those that are proxies for quantities of interest). In most cases this will be a reliable index of abundance related to the spawning biomass (c.f., spawning biomass reference points) or a reliable index of abundance related to the vulnerable biomass (c.f., fishing mortality reference points). The mean length of the composition data associated with the index should also be considered.

Final weights combine the prior weight of overarching hypotheses, weights of the processes not informed by the data or inherently biased, and weights based on hindcast. The probability distributions for the management quantities of interest are calculated for each model (e.g., using a normal approximation), which represents the parameter estimation uncertainty, ensuring they integrate to one, and then combined based on the model weight. The model weights are calculated such that they sum to one to ensure a probability distribution representation. The appropriateness of the normal approximation should be evaluated using posteriors derived from limited MCMC analyses.

## 9. CONCLUSIONS

We have provided a practical guide to applying initial good practices to weight models in tuna risk assessments. This approach is also applicable to other species. However, there is still a substantial amount of research that is needed to improve the good practices. For example, it was concluded at the diagnostic workshop that "current model diagnostics are good for model development, but less so for other purposes" and we would go further to conclude that poor diagnostics may indicate that the model is misspecified but our current understanding is generally insufficient to provide good advice on how the model should be corrected. Substantial additional research is needed to develop guidelines on how to use diagnostics to identify misspecified models, correct the misspecification, and reject models from the ensemble. This will require extensive simulation analyses with a wide range of population dynamics, catch history, and data fit to simulate data which is fit with correctly specified models and incorrectly specified models. A wide range of misspecifications in combination need to be applied. Results from those simulations can then be mined to determine how, in combination, the results of the diagnostics can be used to identify model misspecification and suggest the appropriate correction. Similar simulations are required to determine how hindcasting can be used to weight models. This will provide information on what data should be predicted, what other data should be included in the model, the time periods involved, and how the MASE or other metric can be calibrated to provide an appropriate weight. There is also a contradiction between not using fit to the data to weight models but including parameter uncertainty. Therefore, there is still a need for research to determine the modeling approaches (e.g., data weighting, process variation, among others) to allow the fit to the data to represent uncertainty.

## REFERENCES

Francis, R.I.C.C. 2011. Data weighting in statistical fisheries stock assessment models. Canadian Journal of Fisheries and Aquatic Sciences 68(6): 1124-1138. https://doi.org/10.1139/f2011-025.

Kell, L. T., Kimoto, A., and Kitakado, T., 2016. Evaluation of the prediction skill of stock assessment using hindcasting. Fisheries Research, 183: 119–127.

Kell, L.T., Sharma, R., Kitakado, T., Winker, H., Mosqueira, I., Cardinale, M., and Fu, F. 2021. Validation of stock assessment methods: is it me or my model talking? ICES Journal of Marine Science, 78(6), 2244–2255.

Lee, H-H., Maunder, M.N., Piner, K.R., and Methot, R.D. 2012. Can steepness of the stock-recruitment relationship be estimated in fishery stock assessment models? Fisheries Research 125-126: 254-261.

Legault, C. M., Powers, J. E., and Restrepo, V. R. 2002. Incorporating uncertainty into fishery models. volume Symposium 27, page 208. American Fisheries Society. Section: Mixed Monte Carlo/Bootstrap Approach to Assessing King and Spanish Mackerel in the Atlantic and Gulf of Mexico: Its Evolution and Impact.

Maunder, M.N., Xu, H., Lennert-Cody, C.E., Valero, J.L., Aires-da-Silva, A., Minte-Vera, C. 2020. Implementing reference point-based fishery harvest control rules within a probabilistic framework that considers multiple hypotheses. IATTC Document SAC-11 INF-F REV.

https://www.iattc.org/GetAttachment/46edbd8e-22f9-4bb3-8d26-d4cfd24a472c/SAC-11-INF-F_Implementing-risk-analysis.pdf

Restrepo, V. R., Hoenig, J. M., Powers, J. E., Baird, J. W., and Turner, S. C. 1992. A simple simulation approach to risk and cost analysis, with applications to swordfish and cod fisheries. Fishery Bulletin, 90(4):736 – 748.

SEDAR 2022. SEDAR 78 South Atlantic Spanish Mackerel Stock Assessment Report. SEDAR, North Charleston SC. 177 pp. available online at: http://sedarweb.org/sedar-78

**Appendix 1: Agenda**

**Monday**

8:00 Welcome: IATTC

**Introduction**

8:10 Introduction: Mark Maunder

8:30 Model weighting in the IATTC risk analysis approach: Mark Maunder

9:00 A review of model averaging and the way forward: Carsten Dormann

10:00 Break

10:15 Summary of the ICES 2022 Network Session on ensemble models: Ernesto Jardim

10:45 Discussion

**Tuesday**

**Which models to consider and what measures (diagnostics) should be used to exclude models**

8:00 Using diagnostics to fix and eliminate models when constructing an ensemble: Mark Maunder/Felipe Carvalho

8:30 Discussion

**What measures to use in weighting and how to determine the weight for each metric**

9:00 What measures to use in weighting and how to determine the weight for each metric?: Philipp Neubauer

9:30 Using hindcasting to weight models: Mark Maunder

10:00 Break

10:15 Considerations when integrating multiple stock assessment models: Allan Hicks

10:45 Discussion

**Wednesday**

**How to combine weights**

8:00 Key decisions: How should ensembles be constructed and combined?: Nicholas Ducharme-Barth

9:00 Implementation of Monte-Carlo Bootstrap Ensembles to assess uncertainty and provide management advice: Matt Vincent

9:30 There's more to ensemble modelling than model weighting: Michael Spence

10:00 Break

10:15 Discussion

**Thursday**

**Applications**

8:00 Practical applications of diagnostic weighing in ensemble models: three case studies, Northern shrimp, Adriatic sole and Gulf of Bothnian vendace: Massimiliano Cardinale

**Application of model weighting in the context of EPO tuna assessment and management**

9:00 A proposed approach of developing and weighting an ensemble for use with tropical tunas in the EPO: Carolina Minte-Vera

10:00 Break

10:15 Discussion

**Appendix 2: Abstracts**

**IATTC Risk Analysis**

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

The IATTC staff developed a risk analysis approach to evaluate the probability statements in the harvest control rule for tropical tunas. The risk analysis uses a hierarchical approach to define different hypotheses about the system and each of the hypotheses is represented by alternative assumptions of the stock assessment model. Because a rigorous statistical framework is not applicable to developing weights for stock assessment models due to model misspecification and data weighting issues, a set of multiple metrics were developed to weight the models. The metrics included expert opinion, convergence, fit, plausible parameters, plausible results, and diagnostics. The metrics were scored subjectively by a panel of stock assessment experts. The distributions of the management parameters for each model were combined based on the model weight. The results were presented as probability distributions, cumulative distributions, and decision tables.

**Using diagnostics to fix and eliminate models when constructing an ensemble**

Mark Maunder, Felipe Carvalho, Maia Sosa Kapur, and Andre Punt

Diagnostics have a lot of uses including model selection, model weighting, characterizing uncertainty, data selection, value of information, and stakeholder communication. Here we focus on using diagnostics to create and weight an ensemble of models. Conceptually, diagnostics can be used to determine if a model is adequate and, if not, how to fix it. Eventually, the diagnostics will be used to determine if the model should be used in the ensemble. There are several ways a model can be misspecified including incorrect specification of a fixed model parameter, incorrect model structure, incorrect specification of a likelihood function, incorrect observation model, incorrect systems dynamic model, and ignoring process variability. A variety of diagnostics are needed to detect the different types of misspecification. Some diagnostics are standard to all statistical estimation (e.g., residual analysis, cross validation), others are specific to stock assessment (e.g., R0 likelihood component profile, ASPM). Plausibility of parameter estimates and results can also be used as diagnostics. The CAPAM workshop on diagnostics concluded that Current model diagnostics are good for model development, but less so for other purposes. The development and understanding of diagnostics are not at the stage that diagnostics can be used for weighting models. A comprehensive simulation analysis is needed to better determine the metrics and criteria for each diagnostic.

**Model Weighting using cross validation and hindcasting**

Mark N. Maunder

Hindcasting is a method that can be used to evaluate the reliability of a model through out-of-sample predictions and has been introduced into stock assessment by Laurie Kell and his colleagues. It is based on one-step-ahead predictions due to the time series nature of population dynamics models. However, it can only work on observations and not model estimated management quantities. Therefore, we have to assume that if the observations are predicted well, the model is good, and therefore the estimated management quantities are also good. There are several choices that must be made including what data to predict (e.g., and index of abundance), what data to remove, how many year ahead to predict, and how to measure the reliability of the prediction.

**Considerations when integrating multiple stock assessment models**

Allan C. Hicks, Ian J. Stewart

International Pacific Halibut Commission

Integrating multiple stock assessment models into an ensemble can better characterize uncertainty and provide interannual stability as data are updated. The choice of integration methods is important and depends on exactly how parameters and statistics will be used by managers and others. We highlight some important considerations for dealing with functions of model estimates and provide examples of two approaches to estimate parameters from multimodel stock assessments based on a posterior distribution or proxy that result in two different outcomes regarding management advice along with consequences related to stock status and harvest levels. Furthermore, we consider the number of samples from each model to create statistics that are precise, and discuss some diagnostics, such as running quantiles, that can be helpful when determining the appropriate number of samples. Finally, we present an easy-to-use R package (Ensemble) that can combine MLE outputs from Stock Synthesis models by sampling from the approximate multivariate normal distribution determined from the variance-covariance matrix estimated within AD-Model Builder.

**Implementation of Monte-Carlo Bootstrap Ensembles to assess uncertainty and provide management advice**

Matthew T. Vincent

Stock assessments for Atlantic Ocean and Gulf of Mexico stocks conducted in the Southeast US conducted using the Beaufort Assessment Model, a statistical catch-at-age model, implement a Monte-Carlo Bootstrap Ensemble (MCBE). Recreational fishing is a major source of fishing mortality for several stocks in the region. However, estimates of landings from this sector can be highly uncertain, and sample sizes of age and length frequencies are often moderate to small. A MCBE is conducted to account for uncertainty in the data, fixed parameters, and life history functions used in the assessment. Values of fixed parameters in the assessment, such as discard mortality and natural mortality, are drawn from statistical distributions. Landings and indices are bootstrapped based on coefficients of variations of estimates. Length and age composition data are sampled from multinomial distribution based on the number of samples. Models that meet convergence criteria are retained and their estimates pooled to create distributions of output, including stock status, fishing status, and reference points. Distributions of reference points are used for providing management advice and ensemble models are used to incorporate uncertainty in projections of stock status using proposed management actions. A summary of the strengths of the methodology and areas for improvement is presented.

**There's more to ensemble modelling than model weighting**

Michael A. Spence

Fisheries science, and more generally science is about finding the truth. That could be the true biomass, or abundance, and often we have many sources of direct and indirect information. In principle more information leads to greater, or at worst equivalent, certainty. Different models are different bits of information, and so any methods of combining them should follow this principle. In this presentation I will formally explore how models relate to the truth, showing that ensemble modelling can be generalised to a more familiar class of statistical problems. By examining illustrative examples, I will demonstrate that model weighting is just one way of combining multiple sources of information and that other methods can outperform model weighting schemes. An alternative is to statistically examine the individual model's discrepancy, the difference between the model prediction and the truth, and then use the information from all the models, and their discrepancies, to reduce uncertainty in the truth. Using this statistical meta-model, we can combine prior beliefs, model estimates and direct observations to make coherent predictions with rigorous measures of uncertainty. I will demonstrate this by examining the effects of fishing in the North Sea using four mechanistic multispecies models. I will also introduce EcoEnsemble, an R-package for combining models in this way.

**Practical applications of diagnostic weighing in ensemble models: three case studies, Northern shrimp, Adriatic sole and Gulf of Bothnian vendace**

Max Cardinale, Henning Winker, Francesco Masnadi, David Gilljam and Chris Griffiths

We presented a summary of the results of three newly developed stock assessment ensemble models in European waters for Northern shrimp, Adriatic sole, and Gulf of Bothnian vendace. The models were selected and weighed using a pre-agreed toolbox of diagnostics. The need to weigh models based on information is well recognised in several sciences but it is often difficult to do so within the perspective of fisheries stock assessment models. In this context, model diagnostics have the potential to be used to weight models. Model weighting 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, more unlikely than others.

For all models summarized here, we used a hierarchical approach which included, among others, different data sets, growth curves, natural mortality, predation mortality, steepness and selectivity assumptions. The estimated probability distributions of the quantities of interest for each model were combined (stitched) together using the model weight. The weighting was based on goodness of fit (joint residuals and runs tests), consistency (retrospective analysis), and prediction skill (hindcasting). Models with poor diagnostics were excluded from ensemble (i.e., they were given zero weight).

During the process of model development, it was recognized that the entire process of stock assessment is pervaded by weighting, with excluded models assigned a weight of 0. When weighting, the simple questions we asked to ourselves was: would we prefer a model that can predict the CPUE trend or a model the persistently overestimates the trend? Or a model that is retrospectively stable instead of one that is not? As rarely models pass all diagnostics and model performances might change with time, we preferred to weight than to exclude or equally weight. From a "tactical" perspective, model weights are parameters to be chosen in such a way as to achieve best predictive performance. No specific interpretation of the model is attached to the weights; they must only work.

It was also clear that the weighting scheme must be agreed beforehand to avoid cherry picking of models by the participants, which was highly values by stakeholders present at the meetings. Moreover, a pass/fail system works well for stock assessment models for which diagnostics are used to weigh the models in the ensemble because differences in diagnostics performances between models is often small. Combing (stitching) the distributions of management quantities of each model is important to better represent the tails of the distribution and quantify risk, especially the risk of the SSB to fall below the limit reference point. Finally, simulation testing is needed to determine which diagnostics should be used as weighting metrics.