INTER-AMERICAN TROPICAL TUNA COMMISSION

1st WORKSHOP ON IMPROVING THE RISK ANALYSIS FOR TROPICAL TUNAS IN THE EASTERN PACIFIC OCEAN: MODEL DIAGNOSTICS IN INTEGRATED STOCK ASSESSMENTS

by videoconference
31 January-3 February 2022

CHAIR'S REPORT

Mark N. Maunder, Andre E. Punt, Felipe Carvalho, Henning Winker, Juan Valero, Carolina V. Minte-Vera, and Haikun Xu.

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 Diagnostics in Integrated Stock Assessments on Jan 31-Feb 3, 2022. The workshop focused on defining and automating appropriate diagnostics for fishery stock assessment models and was part of the IATTC work plan to improve the risk analysis for tropical tunas in the Eastern Pacific Ocean (EPO). The workshop was conducted virtually for three hours each day over five days and generally followed the CAPAM workshop style. The workshop is relevant given the diversity of available diagnostics and the lack of agreed good practices for their use. It provided a first step towards addressing calls for the inclusion of diagnostics when conducting integrated assessments. This report is structured around the following three key questions that formed the basis for discussion: (a) which diagnostics are able to identify model misspecification, (b) whether diagnostics can identify what aspect of a model is misspecified, and (c) how an identified misspecification can be addressed. Some diagnostics are based on concepts common across all forms of statistical modelling (e.g. evaluation of residuals and effective sample sizes, hindcasting, and Bayesian model checking). Other diagnostics (e.g. the R0 profile, ASPM and catch curve methods) were developed specifically for fisheries models. Plausibility of parameter values and model outputs could also be part of the process for evaluating models. The use of diagnostics to validate management advice is like extrapolating outside the range of the data. The diagnostics are generally not related to the estimated management quantities (i.e., they relate to the fit to the data, predictions of the data, or other components of the model), and are used to validate the model. It is assumed that if the model is correct, then the resulting management quantities are likely not very biased, but it is unknown whether the quantity the diagnostic focuses on is related to the management quantities of interest. However, there are some exceptions such as hindcasting individual average length and then using it in a management rule (i.e., apply management actions with the goal to make average length equal to that at maturity). In this case, what is predicted is what is being used in management, rather than inferring that when the observations can be predicted, the model is not misspecified. This concept might also work for a management rule based on an index of abundance. 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 and whether the estimates of management quantities are reliable. The development and understanding of diagnostics are 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 hind casting) cannot be turned into P(Model) or made

consistent with AIC. In summary, current model diagnostics are good for model development, but less so for other purposes.

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). Diagnostics are used as part of the scheme to weight models representing alternative hypotheses about the population dynamics and the data used to fit the models. The goal of the IATTC research program is to define more objective, transparent, and automated diagnostic-based metrics for weighting fishery stock assessment model ensembles. The workshop focused on defining and automating appropriate diagnostics for fishery stock assessments, setting the groundwork for the scoring used in the IATTC tropical tuna risk analysis. A future workshop will focus specifically on the scoring.

The workshop was conducted virtually for three hours each day over five days 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. There were 11 invited presentations, 4 contributed presentations, and over 200 participants. The presentations and discussions were recorded and posted on the CAPAM website.
(http://www.capamresearch.org/content/diagnostics-workshop-presentations).
The main deliverable of the workshop is this report.

2. Background

The aim of a system of model diagnostics can be a way to evaluate the estimates of quantities of management interest to corroborate that the model is credible. Diagnostics can be used to select a “best” model, select a set of models to include in an ensemble or to be part of the scheme to weight the models within an ensemble. Diagnostics don’t necessarily have to select models that are correct, but models that are useful (i.e., they provide reliable estimates of the quantities of interest). Diagnostics also are not limited to rejecting or weighting models, but may identify what is misspecified in a model so that it can be communicated and improved. The workshop is relevant given the diversity of diagnostics and the lack of agreed good practices and is a first step towards addressing calls for the adoption of diagnostics when conducting assessments.

The goal of the CAPAM-IATTC workshop was to identify diagnostics that are objective, transparent and can be automated, and to define quantitative criteria for using each diagnostic. The report is structured around the following three key questions that formed the basis for discussions: (a) which diagnostics are able to identify model misspecification, (b) can diagnostics identify what aspect of a model is misspecified, and (c) how can an identified misspecification be addressed. Ultimately, the results of the workshop could be used to develop an “expert system” to evaluate stock assessment performance. However, it was recognized that there will always be an element of subjectivity when evaluating the models on which stock assessments are based. In addition, interpreting diagnostics also needs to consider stock-specific data sampling processes and biology, with particular implications often not directly transferable among assessments. Model misspecification arises from the incorrect specification of (a) parameter values, (b) the system dynamics model, (c) the observation model (including data that are not representative), (d) the likelihood function (sampling distribution), and (f) the sources of process variation.

Some diagnostics are based on concepts common across all forms of statistical modelling (*e.g.* evaluation of residuals and effective sample sizes, hindcasting, and Bayesian model checking). These diagnostics are usually designed to determine if the model assumptions have been violated (*e.g.* the data are not distributed following the distributional assumption assumed in the likelihood function). Other diagnostics (*e.g.* the R_0 profile, ASPM and catch curve methods) were developed specifically for fisheries models. Plausibility of parameter values and model outputs could also be part of the process for evaluating models as used in assessments for tropical tunas conducted by the IATTC (Minte-Vera et al., 2021). However, plausibility is very subjective, and plausibility of parameter values could always be directly incorporated into the assessment as priors.

The order in which diagnostics are applied might be important. For example, applying all the diagnostics to the results of a model that has not converged properly may be meaningless. Some diagnostics may be time consuming (*e.g*. retrospective analysis and hind casting, which require multiple runs of the model), while others are simple statistics calculated from the model results (*e.g.* residual analysis and plausibility of results). If a model is rejected due to easily calculated diagnostics, the more time-consuming diagnostics may not need to be calculated. On the other hand, multiple diagnostics may be needed to diagnose and address model misspecification. The order of diagnostics might be to: 1) conduct convergence tests and if the model fails try methods to achieve convergence before continuing, 2) apply easy-to-calculate diagnostics and if the model fails and the diagnostics identify the problem, fix the model, and finally 3) apply the time- consuming diagnostics.

The use of diagnostics to validate management advice is like extrapolating outside the range of the data (Carvalho et al., 2022). The diagnostics are generally not related to the estimated management quantities (*i.e.,* they related to the fit to the data, predictions of the data, or other components of the model), but are used to validate the model. It is assumed that if the model is correct, the resulting management quantities are not very biased, but it is unknown whether the quantity tested by the diagnostic is related to the management quantities of interest. There are some exceptions. For example, hindcasting average length and then using average length in a management rule (*i.e.*, apply management actions with the goal to make average length equal to the average length at maturity). In this case, what is predicted is what is being used in management, rather than inferring that when the observations can be predicted, the model is not misspecified. Therefore, the estimates of the management quantities are adequate. This concept might also work for a management rule based on the index of abundance. When simulation testing diagnostics, it might be useful to evaluate the diagnostics in terms of how well they diagnose the quantities of management interest rather than the model itself. If the diagnostics are used for weighting models, the weight might be based on the risk related to the bias in the management quantity of interest.

3. Which diagnostics can identify model misspecifications?

Overall, a model would be considered adequate for providing management advice if the optimization was successful, the model fits the data (*e.g*., residual analysis), the model provides reliable estimates of trends and scale, the results of the model are consistent when updated with new data (*e.g.* retrospective analysis), and the model is able to make adequate future predictions (*e.g.* hindcasting) (Carvalho et al., 2021). The following sections outline the primary diagnostics used to evaluate fisheries assessment models, discuss their strengths and weaknesses, identify good practices and highlight areas where future work is needed. Table 1 lists the diagnostic statistics and examples of thresholds that have been proposed for their usage.

3.1. Residual patterns and effective sample sizes

Systematic misfits to the available data should be considered to be a sign of model misspecification and are quantified using residual patterns and how calculated effective sample sizes relate to input sample sizes. These analyses should be interpreted in the context of the model assumptions (e.g. those implied by the likelihood used) and the model assumptions should also be re-evaluated. It may also be important to evaluate residuals with respect to their impact on the management quantities of interest. For example, patterns in residuals for the most recent year may indicate bias for the estimates of current biomass and fishing mortality, while residuals at the start of the time period may be less relevant. Data can be viewed on different levels of aggregation and fits may be good at the aggregated level, but not at the disaggregated level. Consequently, a model may still be useful even if the diagnostics at the disaggregated level are poor because the quantities of interest are typically related to model outputs that are aggregated over time and age/size.

3.1.1 Residual analysis

Most assessment reports include Pearson residuals (for index and composition data). In general residuals are either examined visually or assessed using a runs test or by computing the SDNR[1]. It should be noted that the cutoff values for interpreting runs tests are somewhat arbitrary and different values could be used. Pearson residuals behave poorly for non-normal distributions (*e.g.* Poisson) and alternatives such as probability integrated transformed (PIT) residuals likely perform better. PIT residuals are the same as Dunn-Smyth residuals (Dunn and Smyth, 1996), except that a quantile is randomized between lower and upper bounds when simulating from the probability mass function. Jim Thorson showed, using an example based on a simulated data set, that PIT residuals followed a normal distribution when the model was correct and was non-normal when the model was misspecified. In contrast, Pearson residuals were non-normal even when the model was correctly specified. Use of Pearson residuals could lead to an undesirably high probability of a correctly specified model being rejected for exhibiting residual patterns. Common diagnostics (such as computing the SDNR) should still work with PIT residuals. At present PIT residuals are not computed for common assessment methods such as Stock Synthesis and this should be a research priority.

Traditionally, the bubble plots for composition data have been reviewed visually and this is very subjective (see https://connect.fisheries.noaa.gov/content/23d2b480-3467-4ee2-a93c-17158aa16078
for an interesting hands-on experiment). A more quantitative approach to evaluating bubble plots is needed. Patterns in residuals for composition data could be for: 1) specific age/length or consecutive groups of ages/lengths, suggesting a misspecified selectivity curve, growth model, or other process, 2) a specific year or block of years, suggesting changes in selectivity, growth, or other processes, or 3) a specific cohort, suggesting cohort targeting or cohort-specific growth or other processes. One way to address this is to conduct runs tests over age/length, time, and cohort.

Nicholas Fisch and colleagues suggested the use of the logistic normal (LN) distribution as a way to diagnose model misspecification. In contrast to alternatives such as the Multinomial and Dirichlet-Multinomial distributions, the logistic normal distribution allows for positive correlations among residuals

[1] The SDNR is the standard deviation of the normalized (or standardized) residuals divided by the sampling (or assumed) standard deviation

as well as between-year correlations. These types of residual patterns have been observed for actual assessments and can be generated by, for example, spatial targeting practices. The LN distribution tends to outperform the Multinomial and Dirichlet-Multinomial distribution for simulated data when sample sizes are large but performs quite poorly for small sample sizes. In principle, therefore, a better fit using the LN distribution might suggest model-misspecification. However, further work is needed to prove this. It was noted that a more flexible parameterization of the covariance function when applying the LN distribution may perform better. Factor models could be used to parameterize the covariance function and would shrink (constrain) a maximum likelihood implementation of the LN distribution towards a low-rank variance-covariance matrix. It was also noted that allowing for time-varying selectivity might allow the model to track cohorts/sizes of fish, which could lead to a positive correlation structure, and reduce the need for the LN distribution.

One thing to keep in mind is that the sampling distributions for composition data may not be multinomial. This is particularly important when using simulation methods as diagnostics. The composition data often have outliers and other distortions. This should be taken into consideration when defining likelihoods, conducting data weighting, designing diagnostics, and evaluating residuals.

Andrea Havron outlined six methods for computing residuals in random effects models: (a) full Gaussian, which assumes that the joint distribution of the data and random effects is Gaussian, (b) "OneStepGaussian" where a one-step conditional distribution is approximated using a Gaussian distribution, (c) CDF where the one-step conditional distribution is the ratio of Laplace approximations, applied to log(CDF), (d) MCMC where fixed parameters are mapped to MLEs and *tmbstan* is used to draw a posterior of random effects, I the conditional DHARMa method where new observations are simulated conditional on the fitted random effects, and (f) the unconditional DHARMa method where new observations are simulated given new random effects. The first four methods can be implemented within TMB. Andrea used simulations based on a series of linear models to compare correctly specified and misspecified models. The results suggested that the full Gaussian approach was the fastest, but had the strictest assumptions, the MCMC approach was also fast but tended to be less sensitive to misspecification. The DHARMAa methods were faster than the one-step method and could be used when the CDF of the distribution is not well defined, but could not be used when the random effects were multivariate normal.

One thing to keep in mind when evaluating residuals is that temporal changes in processes are common and these changes are generally not simply random, but show patterns. Patterns in residuals may indicate unmodelled temporal variation in processes. However, they may also be related to temporal variation in the sampling process. One particular type of residual for which it is hard to interpret temporal trends is recruitment deviates from the stock-recruitment relationship. These are different from deviates of the fitted data, but could still be analyzed similarly. Recruitment is expected to show temporal variation due to environmental influences. However, there are examples of applications that estimate substantial regime shifts in recruitment that are suspected to be artifacts of misspecified models (*e.g.* Merino et al., in review).

The workshop participants agreed that these alternative methods for computing residuals show some promise but noted that further validation tests and guidelines were needed, in particular in the context of non-linear models such as integrated stock assessments. Most stock assessments are conducted using penalized likelihood rather than random effects formulations. However, some penalized likelihood problems could be somewhat overcome by implementing marginal likelihood estimation for defining residuals, such as available in TMB (Thorson et al., 2019).

The MCMC method can be used quite generally. It can more appropriately account for random effects, appears to be less sensitive to model misspecification and can test models at various levels. One-step ahead residuals are also attractive, but it is currently not feasible to compute these for Stock Synthesis and similar models. MCMC runs are often conducted as a diagnostic. They can identify local minima, correlations among parameters, asymmetrical uncertainty intervals, etc.

3.2 Likelihood component profile

The likelihood component profile provides a way to identify the influence of information sources on model estimates and a difference in the best estimates of a parameter (or derived quantity) between information sources is suggestive of data conflicts (e.g., Ichinokawa et al., 2014). The R_0 profile is most common, and is used to identify conflicting information in the data about absolute abundance. Component profiles could also be computed for derived outputs such as terminal biomass. Profiles can be based on (penalized) maximum likelihood estimation or Bayesian analysis, but the latter is not common. Although the R_0 profile can identify data conflicts due to model misspecification, the power to detect model misspecification was found to be low by Carvalho et al. (2017).

The recruitment penalty is often the largest component of the R_0 profile and the size of this component can depend critically on whether the recruitment deviations are forced to sum to zero or not. The recruitments are determined by R_0 and the recruitment deviations, so increasing the fixed value of R_0 and decreasing the (average) values of the recruitment deviations can lead to a similar recruitment series and hence fits to data. This manifests itself in the penalty on the recruitment deviates and is why the recruitment deviate penalty if often one of the most influential components of the R_0 profile and is often conflicting with the other most influential components. One approach to counter this is to rescale the recruitment deviates so that they sum to zero (an option in Stock Synthesis). The results are often quite different, but it is not yet clear which the best approach is or what each approach or the differences between them tell us. It might be useful to plot the mean recruitment (as adjusted by the stock-recruitment curve) on the axis rather than R_0. Alternative approaches may include fixing the recruitment deviates at the MLE or doing a profile on the current biomass. Future work should explore how this issue further, although best practice might be to use the option of the profile for which the sum-to-zero constraint is included. In addition, the R_0 profile may change if σ_R is estimated and the model is formulated as a random effects model. Differences between the R_0 profile based on maximum (marginal) likelihood and penalized likelihood should be explored when σ_R is estimated / pre-specified. Whether the steepness of the stock-recruitment relationship should be estimated when conducting a R_0 profile is another question that should be addressed with further research.

Often the profiles are ramps where there is no minimum for a particular component. Often this is because the range of values evaluated is too small, but it also might be that a particular data set only provides information about whether the parameter is too small or too large. In this case, there may not be conflicting information if one data sets ramps to the left while the other ramps to the right, although more research is needed.

3.3 Age-structured production model (ASPM) diagnostic

Running an assessment with pre-specified selectivity and no recruitment deviations, and not fitting the composition data, then checking whether the resulting time-trajectory of spawning biomass is essentially the same as that from the full assessment, is a way to assess whether surplus production and observed catches alone could explain the trend in the index of abundance and hence whether the data (*i.e.*, the indices of abundance) provide information on the scale of the population. Sometimes, the production function can be disguised by a combination of short generation time, recruitment autocorrelation and σ_R. It would be interesting to determine under which life history traits a correctly-specified ASPM would fail to converge in simulations.

There is some confusion regarding the use and interpretation of the ASPM diagnostic. The fact is that the ASPM diagnostic is useful in several situations. For example, when the ASPM-Rdev differs from the fully integrated model it suggests that there is a conflict between the composition data and the index of abundance. When the ASPM differs from the ASPM-Rdev, this indicates that the absolute abundance information in the index and catch cannot be interpreted without information on recruitment variation, which comes from the composition data. For example, a situation in which both the index and the catch goes up (or the catch goes up and the index is relatively flat) can only be explained by increased recruitment (or changes in catchability). Consequently, without the flexibility to vary recruitment, the ASPM will estimate a large biomass so that the catch does not influence the abundance. One possible alternative interpretation is that the CPUE index is not sufficiently standardized to detect the impact of the catch, and the standardization may need improvement.

It seems questionable to weight models by whether the ASPM diagnostic indicates that a production function is evident. However, the ASPM diagnostic should be useful to understand the behavior of a model and whether it is likely able to make reliable forecasts. In addition, the ASPM diagnostic provides information about the relative importance of the index and age-composition data in determining the results of the assessment and it is likely that there are concerns about an assessment for a long-lived species that does not exhibit a production function.

3.4 The catch-curve diagnostic

The catch-curve diagnostic is another possible way to evaluate the information content of the data and understand the behavior of the model. It aims to assess whether the composition data are consistent with the index data or not, which might be suggestive of model misspecification. It can also be used to assess whether the information content of the composition data changes over time. The catch curve diagnostic estimates selectivity and scale parameters, and the model is fit to the composition data (or a subset of composition data). Ideally, if the model is not misspecified, the trends in abundance will be similar to those from a full assessment. However, simulation results in Carvalho et al. (2017) suggest that the catch curve diagnostic performs poorly (high rates of Type I error, i.e. the statistic indicates problems even if the model is not misspecified). Thus, until there is a better understanding of the behavior of the catch curve diagnostic it should be restricted to assessing whether selectivity is correctly formatted during the model development process.

The ASPM and the catch curve diagnostics provide information about the absolute abundance and the conflict between information in indices of relative abundance and composition data. Combining these two diagnostics might provide additional sources of information that would be more valuable than evaluating each diagnostic in isolation. Maunder et al. (2020) provided a flowchart for combining the ASPM and the R_0 profile diagnostics to provide weights for an ensemble model. A similar approach might be useful for combining the ASPM and catch curve diagnostics.

An alternative diagnostic would be to combine the composition data and the index of relative abundance and applying a depletion estimator by cohort (or the method of Clark, 2022). This would involve using the composition data from the index to tune the aggregated index into an index by age and then converting the aggregated catch into total catch-at-age. This could be done using the age-composition data or an approximation by converting the length compositions into age frequencies using the growth equation and associated variation of length-at-age (alternatively using numbers at age and selectivity estimated from an integrated model).

3.5 Empirical selectivity estimation

A package has been developed (R package *empirical.selectivity*[2]) that can compute selectivity based on comparing the composition data (age and length) and estimates of numbers-at-age and at-length. This package can be used to determine the most appropriate form for selectivity by fleet. It also calculates the optimal number of knots for spline-based selectivity patterns.

The package can be useful during final model evaluation as the preliminary model structure is replaced by the final model. In particular, this approach will indicate whether the model-estimated selectivity for older ages/lengths is systematically different from that expected from the data. The older ages/larger lengths are particularly influential when estimating absolute abundance and depletion levels. This comparison is hard to make from standard diagnostic plots (except perhaps bubble plots) given the usually low number of old/large animals in composition data. However, there is currently no way to use the results from the empirical selectivity pattern to weight models.

Standard age/length frequency plots focus on the most abundant ages/lengths in the catch. In contrast, the empirical selectivity method focuses on the least abundant ages/sizes in the population (since the denominator is abundance). The former are usually middle-aged fish while the latter are older fish. The typically used multinomial function weights the fit based on the inverse of the variance *P*(1-*P*), which gives higher weight to smaller proportions and therefore also emphasizes the older fish. It is useful to visualize the residuals by weighting by the inverse of the variance (e.g. bubble plots in r4SS).

3.6 Retrospective analysis

Retrospective analysis allows a comparison of the consistency of model outputs (*e.g.*, spawning biomass, recruitment and fishing mortality and model outputs such as MSY) as additional data are added to an assessment. The results of a retrospective analysis are often summarized using Mohn's rho (Mohn, 1999)[3.] However, the magnitude of retrospective effects can also be quantified using metrics such as the rho-adjustment (Debora, 2014), which approximately corrects for a retrospective pattern. The value of Mohn's rho should be interpreted considering the uncertainty of the assessment, for example based on bootstrapping. Retrospective analysis is usually based on age-aggregated metrics but age-based values for Mohn's rho may be useful for some purposes. The number of peels possible in a retrospective analysis may be limited by data availability, limiting the ability to interpret the results.

Retrospective patterns can be created because the catch time-series is in error (*e.g*. there are missing catches) or some process (*e.g.,* natural mortality, selectivity, or catchability) is time-varying in reality but this variation is ignored (or modelled incorrectly) when conducting the assessment. Retrospective error in the estimates of biomass should not occur if there is an influential index of (adult) abundance as the estimates should always follow the abundance index, except for very short lived species. This may suggest

[2] remotes::install_github("roliveros-ramos/fks")
remotes::install_github("roliveros-ramos/empirical.selectivity")

[3] There are several formulations of the Mohn's rho. A consensus should be developed on which one to use so results can be compared better.

that retrospective analysis primarily detects the influence of the composition data. Similarly, retrospective pattern should not occur in an ASPM unless there are multiple indices that are in conflict.

Often it is possible to remove a retrospective pattern by allowing for time-variation in parameters (either implemented as penalized maximum likelihood or using a random-effects structure). However, it is seldom clear which parameter should be time-varying, and it is likely that multiple processes are actually time-varying. It is possible that a simpler model (e.g. a surplus production model) may outperform a complex model-based assessment that has retrospective patterns. Thus, both adopting more complex and simpler models will not necessarily resolve bias associated with model misspecification. The rho-adjustment (correcting model outputs for retrospective bias) has also been proposed as a way to address retrospective patterns but does not provide an explanation for the cause of the pattern. Evaluating retrospective patterns in more detail, such as for age-specific abundance or fishing mortality, may provide insights into the cause of the problem.

A retrospective pattern indicates that there is systematic error in the model, and it may be fixable. If there is a single large retrospective error, it might be better to ignore it. If there is large, but random retrospective error, it generally means that the model is not very reliable. There are cases where the historical absolute biomass changes as well as recent biomass, but this is not common, and it means that the model is not well formulated and more work is needed to get the scale correct. Care needs to be taken that the retrospective pattern is not fixed by changing the wrong process and the bias increased. Often the retrospective pattern is driven by cohorts disappearing as they age due to *M* and catch not being high enough.

Chris Legault introduced the Rose approach, which involves plotting the results (*e.g.*, current spawning biomass and fishing mortality relative to reference points) for a set of the models that are each constructed to remove the retrospective pattern (Legault, 2020). Example applications of the method suggest that the central tendency of such models leads to results not too different from those obtained through rho adjustment. One alternative suggestion was that when a retrospective pattern exists, the model should be adjusted if the cause of the pattern is known otherwise, the results should be adjusted.

Chris Legault presented several approaches that could be used to quantify the retrospective pattern, including a) evaluating many assessments to determine what the normal Mohn's rho is, b) determining if the adjustment factor is outside the uncertainty estimates, and c) evaluating whether the Mohn's rho uncertainty interval from a parametric bootstrap overlaps zero. The results of actual assessments have been used to assess thresholds for Mohn's rho but it may be better to look at the distribution of Mohn's rho for correctly specified models or at least models with minimal bias in the desired management quantities (c.f. Hurtado et al., 2015).

3.7 Hindcasting cross-validation (HCXval)

Kell et al. (2016) introduced a diagnostic based on using hindcast cross-validation for evaluating prediction skill, defined as the ability to predict actual observations (index of abundance, age-compositions, length-compositions, and tagging data) that have been removed from model. The original work by Kell et al. (2016) explored measures of accuracy of the prediction residuals, such as the RMSE, where a prediction residual is the difference between the observation that is unknown to the model and its model predicted value. A scale free measure of prediction skill is the mean absolute scaled error (MASE) statistic (Kell et al. 2021), for which a value less than 1 indicates performance better than a naive prediction of a random walk (*i.e.* tomorrow's weather will be the same as today). The MASE values can be compared because a value of 0.25, for example, is twice as good as a value of 0.5.

Cross-validation provides a way to evaluate performance for a model or set of models by dividing the data into a training set and a test set. In principle, cross-validation can inform whether there is evidence for overfitting and bias. It can also potentially identify data conflicts and when models should be extended or simplified. Cross validation should be performed based on observations. There are a variety of ways to generate test sets for cross-validating by omitting observations (e.g., sequentially removing entire data years as in a retrospective analysis, all the data for one data series or fleet, the composition data for one entire year). In principle, the Diebold-Mariano test (Diebold and Mariano, 1995) can be used to test the statistical significance of the difference between two sets of forecasts but other measures such as RMSE skill score can be used to evaluate model performance. Input data time-series for stock assessment models are often temporally (or spatially) correlated, so simple random cross validation based on randomly leaving out observations is not appropriate.

Kell et al. (2021) assigned weights to models for Indian Ocean albacore tuna within an ensemble based on AIC weighting, MASE and Mohn's rho. AIC weights are considered less than ideal for this purpose given they depend critically on how each data set is weighted, which is often arbitrary. However, it is not straightforward to use Mohn's rho and MASE to assign continuous weights rather than accept/reject models.

In common with other diagnostics (e.g., residual patterns, retrospective error, R_0 profile) it may be possible to "improve" the hindcasting diagnostics by adding addition processes (*e.g.*, time-varying processes) to the model or dropping data. It is also not clear if the model predictions / estimates of management-related quantities will have been improved. Hindcasting specifically tests if a quantity (*i.e.* data) can be predicted and the uncertainty and bias in that prediction. However, unless that quantity (*e.g.* relative index) is used directly for management advice (*e.g.* empirical harvest control rule), a leap of faith still needs to be made about the reliability of the actual quantity of interest (*e.g.* F/F_{MSY}). In addition, the model cannot predict variation in recruitment or selectivity, which may influence the results, *e.g.* for small pelagic fishes, if present in the application.

It should be kept in mind that even if the model is not able to produce good predictions of one or more data types, this does not mean that the model is not useful for providing management advice. It is also possible that a data set that is problematic to predict is not representative of the population (*e.g.,* only covers a portion of the stock) and therefore a poor prediction may not mean that the model is not good. For example, non-random samples of fisheries-dependent size composition data may be down-weighted to mostly provide information about selectivity (Sharma et al., 2014). By contrast, if the model fails to have prediction skill (MASE < 1) for a scientific survey index that is thought to be an unbiased estimator of the trend in biomass, the ability to make predictions in general is likely low. In general, good MASE performance for CPUE hindcasting is likely to occur if the stock is production driven, the production function is estimable from the data, and the production function is stationary over time. Poor MASE performance is likely caused by violations of one of those three items, which will usually be caused by model misspecification unless the stock is recruitment-driven and predictions of future recruitment are poor.

3.8 Convergence diagnostics

Convergence diagnostics include whether a parameter is on a bound, whether the gradient is small, whether the Hessian matrix is positive definite, whether some of the entries in the correlation matrix are large, and whether a jittering analysis leads to convergence to the same solution for most jitters. None of these diagnostics can guarantee that the parameter estimates correspond to the global minimum of the objective function. Jitter analyses are conducted for many assessments, but care needs to be taken when specifying how much jittering is to be undertaken. Jittering is often conducted after a model is selected, but there may be value in conducting jitter analyses as the first step when conducting an assessment.

ADMB has a new experimental option "-hess_step" which will help with convergence checking. Hess_step refines a near convergent model to full convergence and should crawl along steep ridges in the log-likelihood surface.

3.9 Plausibility

Plausibility is often used qualitatively to rank (or eliminate) models. Plausibility can be based on expert opinion (*a priori* evaluation), whether the parameter values are plausible, and whether the results are plausible.

The nature of the basic data should be considered when assigning plausibility weights to models. For example, the data should be evaluated for representativeness and how data that would be unrepresentative for a model that, for example, was spatially-aggregated are handled should be considered in the qualitative plausibility ranking process.

4. Special considerations for Bayesian assessments

4.1 Fit diagnostics

Most of the diagnostics used to evaluate Bayesian analyses are based on posterior predictive checks. Generically, this involves sampling parameters from a posterior distribution, generating replicated posterior predictive data and comparing metrics of the real data with those based on the simulated data. The comparison can involve visual summaries, but graphs can be subjective.

Posterior predictive checks are a Bayesian approach. In principle, a bootstrap based on the MLE parameter estimates or samples from a multivariate normal distribution based on the variance-covariance matrix for the parameters (with appropriate transformations) could be used to create the parameters vectors on which posterior checks are based. However, evidence suggests that the multivariate normal approximation even with the appropriate transformation may not represent the posterior surface adequately and given the improvements to the algorithms used to sample from posteriors, going fully Bayesian may be best.

Summary statistics can be "omnibus" (*e.g.*, chi-square, Freeman-Tukey, deviance and likelihood ratio) or targeted to the problem at hand (*e.g.*, quantiles of outputs, the proportion of zeros in a data set, and Moran's I for spatial models). Bayesian p-values are often used as summary statistics but they tend to be conservative (i.e. an extreme value is indicative of lack-of-fit but a smallish value may or may not be a problem) so that they may fail to detect a small or moderate lack of fit. Pivotal discrepancy measures (Yuan and Johnson, 2012) based on either parametric or PIT residuals can be used to test lack-of-fit at any scale of a hierarchical model. Cross-validation can also be used to evaluate model fit (and is arguably the gold standard for model evaluation) but can also be computationally very intensive for complex models. A runs test could be used as a targeted discrepancy measure in Bayesian posterior predictive checks. However, it would be necessary to calibrate the discrepancy measure (e.g., Bayesian p value), which requires additional simulations.

There are Bayesian approaches for evaluating model performance but most Bayesian stock assessments have focused on checking whether the MCMC algorithm has converged (or at least there is no evidence that it has not failed to converge).

4. 2 Prior predictive checks

The priors on which Bayesian analyses are based are usually assumed to be independent distributions. However, when parameters are selected from the prior, the resulting model outputs can be *a priori* implausible (e.g., the biomass of an extant stock is less than zero). Prior predictive checks are used to see how credible prior assumptions combined with catches and given the model structure are in terms of model outcomes. Constructing a single joint prior is necessary to ensure plausible ranges of outputs, as parameters may be intrinsically correlated (e.g. due to life-history constraints). Kim and colleagues developed a simulation-based approach that involves (a) taking a sample from a prior distribution, (b) computing the model outputs, (c) identifying those parameters values that do not lead to implausible values, (d) fitting a multivariate normal distribution to the plausible parameter combinations that are then used as prior in step (a), and repeating steps a-d until 99% of the samples lead to plausible model outputs. An undesirable outcome of this process is that the marginal priors are updated even in the absence of data in a likelihood but given the catches and model structure.

Other approaches for achieving a joint prior that leads to plausible model outputs include the use of copulas[4] (*e.g.* Brandon et al., 2007) and reparametrizing the model so that (say) current biomass is a parameter of the model instead of initial biomass.

5. Can diagnostics identify what aspect of a model is misspecified

Once a model fails one or more diagnostics suggesting that it is misspecified, it is useful to determine what component of the model is misspecified so it can be corrected. Some research has shown that it is difficult to determine what components are misspecified because the results of the diagnostics may occur in a data component that is not directly related to the misspecification. For example, misspecification of selectivity may manifest as residuals patterns in the fit to the index data. Currently, there is little evidence to more directly link a diagnostic result to what in a model is misspecified. However, there are some general guidelines about what might be misspecified and trying several fixes to see which of then eliminates the diagnostic issues might be the only reasonable approach. This could lead to several models that are acceptable and an ensemble approach could be used for the assessment. A comprehensive research project is needed to develop the guidelines.

6. How can misspecification be addressed

In general, options to “addressed” model misspecification include dropping one or more conflicting data sources, changing the structure of the model (e.g., by adding additional sources of process error), and down weighting data sources, so the statistical evidence for misspecification is eliminated. Each of these approaches may not perform adequately. For example, “fixing” a model by adding time-variation may not lead to better estimates of management-related quantities even if a retrospective or residual pattern is removed (Szuwalski et al., 2018). An “optimal” approach may be to fix the model misspecification, which may also include re-processing the input data (*e.g.* improving standardizations of index and composition data), but this may not be always possible and there may be multiple misspecifications. In many cases,

[4] A copula is a tool to capture and model dependence structures among variables, by coupling marginal distributions and forming their joint cumulative distribution (Tootoonchi 2021)

adding model complexity or flexibility can improve the diagnostic, but this does not necessarily fix the model misspecification or produce a reliable result. For example, adding flexible time-varying selectivity is an easy way to improve many diagnostics, but it may not necessarily be the appropriate fix. The results of diagnostics need to be linked to the quantities of interest so that improvement in the diagnostic has a high chance of improving the reliability of the quantities of interest.

Estimating the variance parameters of the likelihood functions can also hide model misspecification. The likelihood functions may be specified based on sampling assumptions, but they represent all the variance in the model, including unmodelled process variation and other model misspecification. An "ideal" philosophy (the "leave no variance unexplained" approach) for constructing a model for inclusion in an assessment is to first estimate the sampling variance for each data source based on how the data were collected and to select a model structure so that the residuals have the expected variance given the sampling error (*e.g.*, by allowing for time-varying processes) (Maunder and Piner, 2017; Thorson and Haltuch, 2018). However, it is not straightforward to decide which process should be allowed to be time-varying. Szuwalski et al. (2018) show that allowing any process of natural mortality, growth and selectivity to be time-varying can eliminate residual patterns caused by time-variation in some parameters. Still, the results will be biased if the wrong process is allowed to be time-varying.

An important question is how misfits to the survey age-composition data impact the fit to the survey index of abundance and how misfits to the fishery age-composition impact the catch that is removed from the population. This concerns whether selectivity should be modelled more appropriately or whether simply accounting for the model misfit in the variance parameter of the likelihood function is adequate.

Guidelines are needed to determine when to introduce process variation, in what processes, and the appropriate form (e.g., uncorrelated or autocorrelated), and how to specify or estimate the variance parameter. Including temporal variation in some processes may be a default. For example, it has been suggested that temporal variation in fishery (not survey) selectivity should be a default. There is also the question of what to do when there is insufficient information in the data to estimate temporal variation in some demographic or fishery parameter. Often the variance parameter will go to zero and essentially the model estimates no temporal variation, although often with convergence issues. The temporal variability could be removed from the analysis to improve convergence. On the other hand, temporal variation is likely for all processes. If an informative prior on the variance parameter is available, the temporal variation could be included even in the absence of information, leading to a better representation of uncertainty. However, care needs to be taken when using this approach because adding temporal variation in one process may ignore variation in another process and it is unclear if the bias will be reduced or increased.

In relation to selectivity, it was noted that there is a spectrum between the "VPA approach" where there is no constraint on how much selectivity can vary over time (the model fits the age-composition data very well) and a constrained selectivity parameterization that may be misspecified and lead to severe down weighting of data. The former approach essentially implies that the age-composition data provide virtually no information on trends and/or are known well while the latter implies that the age-composition data are informative and/or imprecisely known but may lead to biased estimates owing to model misspecification. How to decide how much time variation to allow for is still an open question (but see Xu et al., 2019). However, past simulation results suggest that including time-varying selectivity when selectivity is time-invariant is less of a problem than ignoring time-varying selectivity when selectivity is time-varying (*e.g.,* Priviteria-Johnson et al., 2022), suggesting that it is safer to over parameterize (perhaps conditional on the variance parameter being estimated) rather than adopt a model with a highly parametric formulation. Methods that treat both the amount of temporal variation in selectivity and the variance of the data as estimable parameters may help determine the compromise between the two.

One suggestion was to allow for time-variation in the extent of time-variation of selectivity (i.e., time-variation on the variance of the selectivity deviations), with this variance greater when known changes in the fishery occurred. Jim Thorson is working on an extension of the Dirichlet-Multinomial likelihood (derived from a marked log-Gaussian Cox process and implemented using a multivariate Tweedie) that estimates over dispersion as well as heteroscedasticity. Further simulation work to better understand best practices for selectivity estimation is warranted.

6.1 Getting the likelihood function right

There are two main reasons to ensure that the correct data weighting (i.e., variance parameter values) is used. The first is to ensure that the overall data weighting is correct so that estimates of total uncertainty (e.g., which influence confidence intervals, model selection, and model weighting) are appropriate. The second is to ensure that each data source is assigned the appropriate relative influence. The likelihood function measures the total variance of the model fit to the data and therefore includes the random sampling error, model misspecification, and unmodelled process variation. If the variance parameter is estimated, it will reflect these three sources of variation. However, the model may be misspecified and the results biased.

The three sources of variation could be separated. The variance of the random sampling error could be estimated from the data and fixed in the likelihood. The caveat here is that the data should be representative of the stock (i.e., the observation model should be correct). For example, if there is spatial structure and the data are only from one area, then the observation model must connect those data to the appropriate spatial component of the population and the process model needs to model the spatial structure.

Additional process variation could be modelled either by using a state-space model or by modelling random effects on individual processes. Simple state-space models often include all the process variation on a single error term on the abundance (by age) and typically assume that the temporal variation is the same for all time periods (and sometimes ages). However, the importance of different processes (e.g., fishing versus natural mortality) might change over time (and among ages) and it may be important to explicitly model the temporal variation for each process. Temporal variation in specific processes (e.g., fishing mortality or selectivity) modelled as a random effect can be combined within a state-space model. Modelling process variation when it does not exist only likely slightly increases the variation in model outputs, but not modelling temporal variation when it does exist, particularly if there are trends or shifts, can lead to large bias. It has been shown (e.g., Thorson et al. 2019) that stock structure and temporal variation in the spatial distribution of the fishery results in time-varying selectivity and is likely to occur for all fisheries. This implies that fishery selectivity should always be modelled as time-varying but may require the index to have time-invariant selectivity. However, modelling temporal variation in one process (e.g., selectivity) when the temporal variation actually exists in a different parameter can also lead to bias.

Estimation of the variance parameter for the temporal variation can be problematic. The often-used penalized likelihood approach results in a likelihood that is degenerative towards zero with the possibility of a negatively biased local optimum. It is more appropriate to use a marginal likelihood that integrates across the process variation, but this can be computationally intensive, result in convergence issues, and the resulting estimates of the variance parameters may be confounded with the estimates of the variance parameters for the likelihood function (*i.e.*, presumed extent of observation error).

Likelihood functions often ignore the correlation in residuals. This correlation may be due to the sampling process or model misspecification. Most likelihood functions used for composition data are based on the multinomial variance structure that has negative correlations due to the totals of all the proportions summing to one. Use of these likelihoods may result in effective sample sizes and weighting that are too large. The logistic-normal has been proposed to model the correlations (e.g., through an AR1 process), but the error is often higher than multinomial unless sample sizes are high. This is because additional parameters are estimated increasing the variance. Even when multinomial variance is used, and the variance parameter is estimated (*e.g.*, using the Dirichlet likelihood), it has been shown that it may be better to fix the variance parameter if it is known within half an order of magnitude (Maunder, 2011).

Evidence for correlated residuals, which can be evaluated through the difference in effective sample sizes calculated by the Dirichlet versus the logistic normal, could indicate model misspecification. To separate the correlations in residuals caused by the sampling from those caused by model misspecification, the sampling component could be calculated from the data and fixed in the likelihood function. Then the model misspecification component could be minimized by fixing any model misspecification and the remaining correlation either ignored or estimated.

7. Other

Diagnostics are currently used in a very interactive fashion and often evaluated subjectively. Automation and reducing subjectivity has some advantages, but without the user interaction there is the risk that automation may allow models with issues that would otherwise be identified by the user to be missed.

8. General discussion

The diagnostics discussed in the workshop focused on fitting stock assessment models to data. There was no evaluation of the data that goes into the models. The observation model and the sampling distribution assumptions (e.g., the likelihood function) that are used to relate the data to the model need to be appropriate. In general, this means that the data have to be representative of the population. For example, a survey may not be representative of a population if it only covers a small portion of the population's habitat, and the results of the survey should not be included in the assessment unless the observation model can correctly account for the fact that the survey only represents a small portion of the habitat.

It is clear that there has been substantial progress on developing new and modifying existing diagnostics for fisheries stock assessment. However, much more research is needed to develop a set of reliable diagnostics and their evaluation criteria that can be used to evaluate and improve fishery stock assessment models. Winker presented a way forward by conducting a comprehensive simulation analysis using a variety of operating models and assessment models. By comparing the results from candidate diagnostics from correctly specified models and models with various misspecifications (possibly in combination), the results could be used to determine which diagnostics and results identify what model misspecification. This analysis could also be used to develop the quantitative criteria for each diagnostic. It could also be used to determine the probability of detecting a model misspecification when it is not present. General simulations might be useful for guidelines, but simulations specific to each application might also be needed. This is a large task and a coordinated collective project might be needed to produce the appropriate guidelines.

Diagnostics have been used to reject models, identify model misspecification and fix models, and weight models. A system should be developed to clearly outline how diagnostics should be used when providing management advice. One approach is to use diagnostics to initially fix models and then eliminate the models that cannot be fixed. This approach is illustrated in Figure 1. First, a set of candidate models are chosen based on the assessment authors' knowledge and experience (*e.g.* based on a conceptual model for the dynamics of the stock). Then diagnostics are applied to each of these models. Those that pass all the diagnostics go through to the final model ensemble. Those that fail are modified as indicated by the diagnostics. This might mean each original model is modified in multiple ways producing multiple models, which are then tested using the diagnostics. If one of these models passes all of the diagnostics, it is then included in the model ensemble. This is repeated with the remaining models until no fixes are suggested by the diagnostics. The models in the ensemble can then be weighted by their fit to the data (assuming the data weighting is adequately addressed).

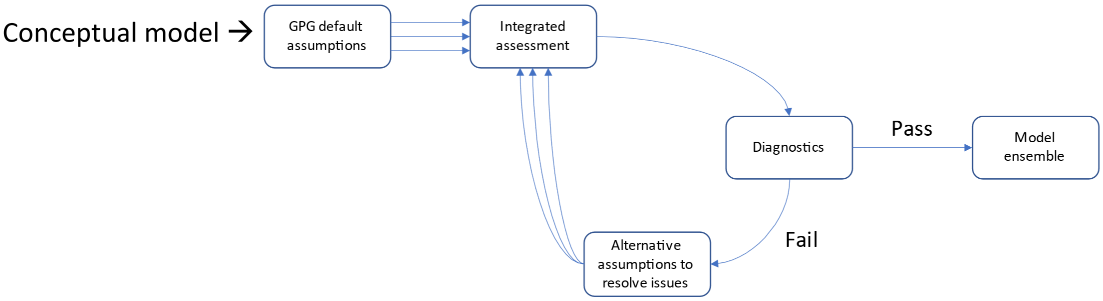


FIGURE 1. Example of an expert system to construct an ensemble of models for fisheries stock assessment.

If diagnostics are to be used to weight models and the results of the assessment, they have to be translate into P(Model). This requires the test metrics to be calibrated and this may only be possible through extensive simulations as described above.

Model structure is often based on achieving "parsimony". This is a compromise between low model complexity and the ability to fit the data and is generally related to the ability to predict observations not used to fit the model and represents a bias-variance tradeoff. A simple model has biased predictions, while a complex model has variable predictions. Unfortunately, the quantities of interest from a stock assessment model are not the observations used to fit the model. So, selecting a parsimonious model in the traditional sense is not appropriate. Traditionally, linear models or models with parameters fixed at zero would be tested against nonlinear models or models with their parameters estimated. However, since the quantity of interest is not the data, using a parsimonious model that predicts the data may not be appropriate. Also, linear models may not be the right parsimonious model. For example, take the Schaefer production model that is linear in some respects and is often used as a default. The Schaefer model has B_{MSY}/B_0 occurring at 0.5, while most age-structured assessments have B_{MSY}/B_0 occurring at less than 0.5. Therefore, it does not make sense to use the Schaefer model even if it is chosen over a Pela-Tomlinson (PT) model, which has one more parameter (the shape parameter of the PT model can be fixed at a value that represents the Schaefer model). Instead, a Pela-Tomlinson model with the shape parameter fixed at an appropriate value or a fully age-structured model with the parameters representing growth, natural mortality, the stock-recruitment relationship, and selectivity fixed at appropriate values, should be used. An intermediate complexity model would be a Bayesian model with informative priors on the shape parameter.

Diagnostics used for data-poor stocks may differ and potentially be less useful than those used for data-rich stocks. Similar standards (of diagnostic procedures/cookbooks) need to be developed for data-poor assessments. Ultimately, diagnostics should be clear and transparent when transmitting advice about stock status for data-poor stocks and should acknowledge that the advice will be inherently more uncertain than the advice produced for data-rich stocks. An alternative approach is to use data-rich approaches with all the associated diagnostics in data-poor situations by making explicit assumptions about model parameters and structure.

9. Conclusions and recommendations

- 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.
- 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 hind casting) cannot be turned into P(Model) or made consistent with AIC.
- The observations are not the quantities of interest. Therefore, we are essentially extrapolating outside the range of the data to latent quantities for providing management advice. This implies that our model assumptions must be approximately correct for the extrapolation to be useful. Hindcast cross-validation of data that are unknown to the model is useful for validating the model's prediction skill of trends abundance, but this approach can still not guarantee that the model correctly classifies the stock status and that the management advice is reliable.
- The analyses that led to rules of thumb for retrospective patterns (Hurtado et al., 2015) should be repeated given newer assessment methods.
- PIT residuals and hindcast cross-validation should be included in diagnostic packages and an investigation conducted as to whether they are better at detecting model misspecification.
- Methods should be developed to detect cohort-specific residual patterns, perhaps using a selectivity formulation that allows for cohort- as well as age- and year-specific random deviations.
- Current model diagnostics are good for model development, but less so for other purposes.
- Alternative validation-based metrics should be explored that are suitable for model weighting by being consistent with an AIC or likelihood, e.g., a "prediction likelihood" that can be computed based on prediction errors from hindcast cross-validation (c.f., Dormann et al., 2018)

REFERENCES

Brandon, J.R., Breiwick, J.M., Punt. A.E., Wade, P.R. 2007. Constructing a coherent joint prior while respecting biological bounds: application to marine mammal stock assessments. ICES J. Mar. Sci. 64, 1085-1100.

Carvalho, F., Winker, H., Courtney, D., Kapur, M., Kell, L., Cardinale, M., Schirripa, M. *et al.* 2021. A cookbook for using model diagnostics in integrated stock assessments. Fish. Res. 240, 105959.

Clark, W.G. 2022. Why natural mortality is estimable, in theory if not in practice, in a data-rich stock assessment. Fish. Res. 248, 106203.

Deroba, J.J. 2014. Evaluating the consequences of adjusting fish stock assessment estimates of biomass for retrospective patterns using Mohn's rho. N. Am. J. Fish. Manage. 34, 380–390.

Diebold, F.X., Mariano, R.S. 1995. Comparing predictive accuracy. J. Bus. Econ. Stat. 13, 253-263.

Dormann, C.F., Calabrese, J.M., Guillera-Arroita, G., Matechou, E., Bahn, V., et al. 2018. Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. Ecol. Monogr. 88, 485–504.

Dunn, P.K. Smyth, G.K. 1996. Randomized quantile residuals. J. Comp. Graph. Stat. 5, 236-244.

Kell, L.T., Kimoto, A., Kitakado, T. 2016. Evaluation of the prediction skill of stock assessment using hindcasting Fish. Res. 183, 119-12.

Kell, L.T, Sharma, R., Kitakado, T. Winker, H., Mosqueira, I., Cardinale, M., Fu, D. 2021. Validation of stock assessment methods: Is it me or my model talking? ICES J. Mar. Sci. 78, 2244-2255.

Hordyk, A.R., Huynh, Q.C., Carruthers, T.R. 2019. Misspecification in stock assessments: Common uncertainties and asymmetric risks. Fish Fish. 20, 888–902.

Hurtado-Ferro, F., Szuwalski, C. S., Valero, J. L., Anderson, S. C., Cunningham, C. J., Johnson, K. F., Licandeo, et al. 2015. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. ICES J. Mar. Sci. 72, 99–110.

Maunder, M.N. 2011. Review and evaluation of likelihood functions for composition data in stock-assessment models: Estimating the effective sample size. Fish. Res. 109, 311–319.

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/Meetings/Meetings2020/SAC-11/Docs/_English/SAC-11-INF-F_Implementing%20risk%20analysis.pdf

Minte-Vera, C., Maunder, M.N., Aires-da-Silva,, A.M. 2021. Auxiliary diagnostic analyses used to detect model misspecification and highlight potential solutions in stock assessments: application to yellowfin tuna in the eastern Pacific Ocean. ICES J. Mar. Sci. 78, 351-3537.

Mohn, R. 1999. The retrospective problem in sequential population analysis: An investigation using cod fishery and simulated data. ICES J. Mar. Sci. 56, 473–488.

Privitera-Johnson, K., Methot, R.D., Punt, A.E. 2022. Towards best practice for specifying selectivity in age-structured integrated stock assessments. Fish Res. 149, 106247

Sharma, R., Langley, A., Herrera, M., Geehan, J., Hyun, S.Y., 2014. Investigating the influence of length-frequency data on the stock assessment of Indian Ocean bigeye tuna. Fish. Res. 158, 50–62.

Szuwalski, C.S., Ianelli, J.N., Punt, A.E., 2018. Reducing retrospective patterns in stock assessment and impacts on management performance. ICES J. Mar. Sci. 75, 596–609.

Thorson, J.T., Haltuch, M.A. 2018. Spatio-temporal analysis of compositional data: increased precision and improved workflow using model-based inputs to stock assessment. Can. J. Fish. Aquat. Sci. 76, 401–414.

Thorson, J.T., Rudd, M.B., Winker, H., 2019. The case for estimating recruitment variation in data-moderate and data-poor age-structured models. Fish. Res. 217, 87–97.

Xu, H., Thorson, J.T., Methot, R.D., Taylor, I.G., 2019. A new semi-parametric method for autocorrelated age- and time-varying selectivity in age-structured assessment models. Can. J. Fish. Aquat. Sci. 76, 268–285.

TABLE 1. Summary of characteristics of the diagnostics. [Partially automated means that it can be automated for a particular application, but is complicated to automate in general]

Diagnostics	**Quantitative criteria**	**Automated**	**Should be used to help diagnose model misspecification**	**Select models to include in ensemble**	**Weight models**
Residual analysis	Runs test	Yes	Yes	Yes	Potential
R_0 profile	No	Yes	Yes	Yes	No
ASPM	No	Partially	Yes	Yes	No
Catch Curve	No	Partially	Yes	Yes	No
Empirical selectivity	No	Potential	Yes	Yes	No
Retrospective analysis	Mohn’s Rho	Yes	Yes	Yes	Potential
Hind casting	MASE	Yes	Yes	Yes	Potential

Appendix 1: Agenda

Tentative Agenda

Jan 31-Feb 3 9am-1pm (San Diego time).

Monday

9:00 Welcome and instructions, Alex Aires-da-Silva and Mark Maunder

9:10 Introduction, Mark Maunder

9:30 The value of diagnostics in stock assessment, Felipe Carvalho

10:30 “and he saith unto them, Follow me, and I will make you fishers of [data]” (i.e., how to sort and weigh data). James Thorson.

11:00 Break

11:15 The Logistic-normal as a tool to diagnose model misspecification? The proposed idea, its comparison to common diagnostics, and some initial considerations. Nicholas Fisch

11:45 Guidelines to validating generalized linear mixed models in Template Model Builder using quantile residuals. Andrea Havron

12:15 Discussion

Tuesday

9:00 Retrospective analysis as a diagnostic. Chris Legault

9:30 There is a crack in everything, that’s how the light gets in. Hindcasting for model validation and selection. Laurie Kell.

10:00 The Art of Bayesian Model Checking. Paul Conn

10:30 Importance of prior predictive checks in Bayesian stock assessment models. Kyuhan Kim

11:00 Break

11:15 Jim Ianelli and Paul Spencer. Use of posterior predictive intervals in complex statistical agestructured assessment models

11:45 Discussion

Wednesday

9:00 R0 likelihood component profile as a diagnostic tool: thumbs up or thumbs down. Hui-Hua Lee

9:30 Age-structured production model and catch curve analysis diagnostics for integrated models. Carolina Minte-Vera

10:00 Empirical selectivity. Carolina Minte-Vera

10:30 Discussion

11:00 Break

11:15 Diagnostics in Stock Synthesis. Ian Taylor

11:45 A guide to using ss3diags for model evaluation, Megumi Oshima

12:15 Discussion

Thursday

9:00 On automating assessment model diagnostics and the need for simulation testing, Henning Winker

10:00 Discussion

10:30 Break

10:45 Diagnostics, yesterday, today and tomorrow. Andre Punt

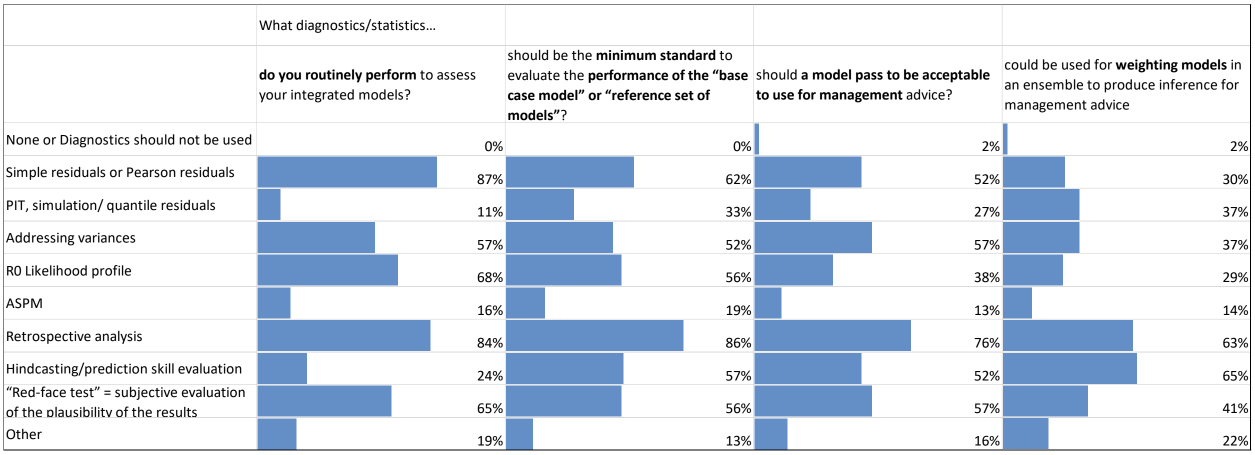
11:45 Discussion

Appendix 2: Online survey results

A survey was conducted among the participants of the workshop regarding the use of diagnostics (46% participation). All respondents routinely perform diagnostic to assess their models (Figure 1). The most used diagnostic was simple residuals or Pearson residuals (87%) and retrospective analyses (84%). About 2/3 of respondents also use subjective evaluations of the plausibility of the result, compute R0 profiles and address variances (e.g. evaluate if the residual variance of the fit is consistent with that assumed in the likelihood function). The age-structure production model, hindcasts and simulation residuals are not yet conducted routinely.

The respondents consider that minimum standards to evaluate models should include at least retrospective analysis (86% of responses, Figure 1), but ideally should include simple/Pearson residual analysis, R0 profile, hindcasting/prediction skill evaluations, as well as addressing variances and include subjective evaluations of the plausibility of the models. More than 50% of respondents considered that a model should pass all those diagnostics to be acceptable for use for management advice.

Many stock assessments use the ensemble model approach to address uncertainty rather than the "best-case model" approach. The management advice and results are presented as a combination of the results from multiple models. It is not straightforward to weight the result of each model, although those weights are key to combine the results in a model ensemble. Several methods have been proposed, and the use of diagnostics in combination with other information is one possible approach. Almost all respondents think diagnostics should be used to weight models (Figure 1). Retrospective analysis and hindcast/prediction skill evaluation were cited by more than 60% of respondents as diagnostics that could be used for that purpose. There is no consensus on other diagnostics to use, as all other diagnostics were cited by about 30% of respondents, except for age-structure production model that was cited by 14% only.



	What diagnostics/statistics…			
	do you routinely perform to assess your integrated models?	should be the **minimum standard** to evaluate the **performance of the "base case model" or "reference set of models"**?	should **a model pass to be acceptable to use for management** advice?	could be used for **weighting models** in an ensemble to produce inference for management advice
None or Diagnostics should not be used	0%	0%	2%	2%
Simple residuals or Pearson residuals	87%	62%	52%	30%
PIT, simulation/ quantile residuals	11%	33%	27%	37%
Addressing variances	57%	52%	57%	37%
R0 Likelihood profile	68%	56%	38%	29%
ASPM	16%	19%	13%	14%
Retrospective analysis	84%	86%	76%	63%
Hindcasting/prediction skill evaluation	24%	57%	52%	65%
"Red-face test" = subjective evaluation of the plausibility of the results	65%	56%	57%	41%
Other	19%	13%	16%	22%

FIGURE 1. Results of a survey conducted using the Zoom platform with the CAPAM diagnostics workshop participants (46% participation).

Appendix 3: Abstracts

Inter-American Tropical Tuna Commission (IATTC) and the Center for the Advancement of Population Assessment Methodology

Mark N. Maunder

The workshop on "Model Diagnostics in Integrated Stock Assessments" was motivated by IATTC's risk analysis approach for tropical tunas, which used diagnostics as a weighting system. Diagnostics were used to weight the model, rather than standard statistical methods (e.g. AIC), because a rigorous statistical framework is not applicable since stock assessment models are complex and highly parameterized, they are misspecified, process variation is ignored, and data are not weighted appropriately. However, the weighting approach was based on the subjective judgement of a panel of experts who had differences in interpretation of the diagnostics and the resulting scoring. Therefore, the goal of the workshop is to design an approach that is more objective, transparent, and automated.

CAPAM's purpose is to use our combined knowledge to create a "Good Practices Guide" (GPG) for stock assessment. The GPG can be used to determine default assumptions for developing a stock assessment. However, the assessment will likely need fine tuning for the specific application. Diagnostics can help ensure the model assumptions are met and provide possible solutions if they are not. The process may involve developing an expert system to construct an ensemble of models for fisheries stock assessment. This would involve evaluating several alternative models using diagnostics, fixing them if possible, eliminating the bad ones, and retaining the good ones. Using diagnostics requires modification of the "Law of Conflicting Data" to include conflict between the model and data to ensure that the diagnostics are interpreted in the context of the random sampling error.

The Value of Diagnostics in Stock Assessment

Felipe Carvalho and Henning winker.

The goal of this presentation is first to provide reasoning on why we should consider apply model diagnostics throughout the assessments development process. First, a brief review on the implementation and interpretation of some current model diagnostic tools is provided. Followed by description of recent innovations developed to help the application and improvement of model diagnostics, and what steps fisheries organizations are taking to encourage the application of diagnostics on their routine assessments.

“and he saith unto them, Follow me, and I will make you fishers of [data]” (i.e., how to sort and weigh data)

James Thorson

Survey scientists often conduct bulk sampling and then stratified subsampling, e.g., where a bottom trawl yields samples for hundreds and species, and then individual specimens are subsampled and individually measured. Similarly, stock assessment scientists typically receive a dump of unsorted data, and must weight each datum to specify its leverage upon estimated parameters. Assessment scientists inspect model fit and residual diagnostics to assign data weights (and to discard data that are erroneous or suspect), and there are fewer standards for sorting data than for survey deck samples.

I here provide three suggestions for practical data-weighting standards:

1. I recommend replacing Pearson- with simulation-residuals using a probability integral transform (PIT). In support, I use the empirical distribution for a Poisson distribution to show that Pearson residuals generate more positive outliers than expected, and a catch-curve simulation to show that PIT residuals can distinguish a correctly and incorrectly specified model when Pearson residuals cannot.
2. I recommend distinguishing the sample size of expanded data (“input sample size”) and the sample size of residuals (“effective sample size”), and using their ratio to diagnose model components that require further elaboration. In support, I review available estimators for both tasks as well as evidence that this approach can diagnose and correct-for misspecified selectivity.
3. I recommend making both deck-sorting and sampling weighting more efficient and less idiosyncratic. In support, I review recent research that can track this pipeline from deck-sorted animals (“nominal sample size”) through to the variance of assessment-model output. This pipeline involves separate analyses, e.g., a bootstrap simulation for nominal to input sample size, and an analytic estimator for input to effective sample size.

Given the small proportion of fished stocks that have quantitative population models in the US and worldwide, I hope that efficiency improvements will support benefits to ocean management via continued growth of rigorous stock assessments.

The Logistic-normal as a tool to diagnose model misspecification? The proposed idea, its comparison to common diagnostics, and some initial considerations

Nicholas Fisch, Ed Camp, Kyle Shertzer, Robert Ahrens, and Mark Maunder

Self-weighting likelihoods, those that estimate the expected variance in residuals within an integrated model, have increased in use since the incorporation of the Dirichlet-multinomial into Stock Synthesis and other popular software packages. This stems from the ability of the likelihood to account for both increased observation error and some degree of model misspecification by down-weighting, or increasing the expected variance of the residuals for the fit to a composition data set. Recent research has suggested that the Logistic-normal performs better than the Dirichlet-multinomial under two conditions: 1) large samples size for composition data and 2) a large degree of model misspecification. With little misspecification and a large enough sample size, the likelihoods were shown to perform similarly. Thus, it follows that differences between a model fit with the Logistic-normal and one fit with the Dirichlet-multinomial, conditional on an adequately large sample size, would suggest a non-trivial degree of model misspecification. Such differences may then serve as a diagnostic to identify model misspecification. The theory is as follows: The Logistic-normal, through its ability to specify a variance-covariance matrix different to that of the multinomial, is better able to account for variability and correlations in residuals as a function of model misspecification than is the Dirichlet-multinomial. Conditional on an adequate sample size, differences between a model fit with the Dirichlet-multinomial and the Logistic normal then suggest misspecification in the model. In this presentation we propose this idea in the context of other model diagnostics common for integrated assessments, specifically with respect to two stock assessments at opposite ends of the spectrum with respect to the number of fish sampled for composition data: Cobia and Pacific Hake. We will then offer considerations, cautions, and recommendations for future research in relation to this proposal.

Guidelines to validating generalized linear mixed models in Template Model Builder using quantile residuals.

Andrea M. Havron1, Cole Monnahan2, Florian Hartig3, Kasper Kristensen4, James T. Thorson2

Model validation, whether via graphical examination or formal hypothesis tests, is a crucial step in statistical analyses to alert the analyst to potential model misfit. Unlike model selection, where models are compared relative to each other, model validation looks for inadequacies in a single model and helps identify likely causes and thus potential solutions. Generalized linear mixed models (GLMMs) are now ubiquitous in the ecological literature, yet well-established protocols for model validation are lacking compared to generalized linear models. This is partly because Pearson residuals are unreliable for model validation, and the more appropriate quantile residuals are more difficult to calculate and interpret.

Here, we review two distinct statistical approaches for quantile residuals: one-step-ahead (OSA) and simulation residuals via DHARMa, both of which are readily available in the popular TMB package. We review the statistical properties, interpretation, and calculation of each method and their variants, including their application in TMB. We test the calibration of p-values from normality tests of residuals using simulation on three examples: a GLMM, time series model, and spatial model. We also provide examples of several different types of mis-specifications to demonstrate appropriate model validation statistics and test whether the mis-specified models fail validation. In certain cases, simulation residuals were faster to calculate and had good performance, but OSA residuals were better in the multivariate context. Overall, the flexibility of generalized linear mixed models can affect the power needed to detect model mis-specification.

Retrospective analysis as a diagnostic

Christopher M. Legault

Retrospective analysis examines the stability of a stock assessment by sequentially removing the most recent year of data and re-running the model. Strong, directional changes in estimates of biomass or fishing mortality indicate something has changed in the data or model assumptions that does not match the model formulation. Identifying the source of a retrospective pattern has been largely unsuccessful because multiple changes result in similar patterns. Strong retrospective patterns lead to poor management advice if left unaddressed and so have been used as the basis to reject stock assessment model formulations. In such a situation, the stock assessment can be modified in different ways. The model can rho-adjusted before providing management advice, but this leads to an inconsistent time series. Reverting to simpler models that use less data and often do not exhibit retrospective patterns does not necessarily provide improved management advice. State-space models can reduce retrospective patterns relative to statistical catch-at-age models due to their increased flexibility, but can still exhibit retrospective patterns. Fixing the model with the incorrect source of the retrospective can lead to poor management advice. The Rose is a new ensemble approach that uses multiple fixes and averages across them all in acknowledgement of the difficulty in discerning the true source(s) of the retrospective pattern, but is time consuming. More research is needed on statistical properties of retrospective patterns under a wide range of situations to move retrospective analysis beyond being just a diagnostic to providing guidance on how to improve the stock assessment and management advice.

There is a crack in everything, that's how the light gets in. Hindcasting for model validation and selection.

Laurence T. Kell, Massimiliano Cardinale, Henning Winker, Iago Mosqueira, Rishi Sharma, Toshihide Kitakado

There two ways to conduct a hindcast based either on observations, i.e. crossvalidation, or model estimates, e.g. as a backtest. While there are three reasons for doing so, namely to find the "best assessment", select and weight models in an ensemble, or condition Operating and Observation Error Models when conducting Management Strategy Evaluation. We review how stock assessment models are currently validated, summarise the use of the hindcast by the RMFOs, and propose how to adopt hindcasting as an objective approach for selecting, screening and weighting hypotheses.

The Art of Bayesian Model Checking

Paul Conn

In this talk, I introduce the basics of Bayesian model checking to an applied audience. As this literature is quite large, I focus on techniques and procedures that are likely to have high practical value. This necessarily includes a discussion of discrepancy functions and posterior predictive checks, but also less well known approaches such as sampled posterior p-values and pivotal discrepancy measures. I also emphasize the importance of probability integral transformations as a flexible approach for evaluating fit at various levels of hierarchical models. I demonstrate these approaches when assessing the fit of a Bayesian spatial regression model. Finally, at a request from the workshop organizer, I provide a brief overview of a goodness-of-fit procedure developed for integrated population models. The presentation and underlying R Markdown code will be made available at www.github.com/pconn/BMC_CAPAM_talk.

Importance of prior predictive checks in Bayesian stock assessment models

Kyuhan Kim, Philipp Neubauer, and Kath Large

Fisheries management decisions in New Zealand are based on stock assessment models where inferences are typically carried out using a Bayesian approach. In any Bayesian model, the choice of priors on model parameters is a key part. However, when introducing this prior knowledge into a stock assessment model, we tend to focus on the biological meanings of individual parameters (e.g., log-normal distributions on positive parameters, such as the intrinsic growth rate, r, and the carrying capacity, K, in surplus production models), rather than the joint prior distribution of parameters in the context of the model. Using simulation studies of the two most common stock assessment models, logistic production and age-structured models, we demonstrate that assuming independent priors, where each marginal prior contains true information on individual parameters, can potentially drive stock assessment models to a priori implausible spaces, leading to biased inferences for key model parameters. We further show that considering a joint prior in terms of plausible ranges of model outcomes (e.g., non-negative values for stock biomass) is a necessary step in Bayesian stock assessments in order to eliminate this bias. In statistics community, it is already well recognised that a poor choice of priors in terms of model outcomes can cause underpowered inferences (Kennedy et al., 2019), and priors should only be interpreted in terms of the likelihood (Gelman et al., 2017). In this talk, we show that Bayesian stock assessment models are no exceptions to the necessity of prior predictive checks, and we suggest a procedure which extends stochastic stock reduction analysis to combine domain knowledge into a joint prior that minimises prior conflict and bias in posterior inferences.

References

Gelman, A., Simpson, D., Betancourt, M., 2017. The prior can often only be understood in the context of the likelihood. Entropy. 19, 555. https://doi.org/10.3390/e19100555

Kennedy, L., Simpson, D., Gelman, A., 2019. The Experiment is just as Important as the Likelihood in Understanding the Prior: a Cautionary Note on Robust Cognitive Modeling.Comput Brain Behav. 2, 210–217. https://doi.org/10.1007/s42113-019-00051-0

Use of posterior predictive intervals in complex statistical age-structured assessment models

Jim Ianelli and Paul Spencer

Model fitting diagnostics that describe exceptional circumstances are useful to determine the utility of an assessment for catch advice. Here we demonstrate a way to graphically present model fits using posterior predictive intervals from a case study for Eastern Bering Sea walleye pollock. This approach was developed with a view for improving model selection procedures when environmental covariates are evaluated. We assert that this approach can help bring models of intermediate complexity (e.g., multispecies) into tactical management advice.

Age-structured production model, catch curve analysis and empirical selectivity diagnostics for integrated models

Carolina Minte-Vera

In this presentation the age-structure production model and the catch-curve analysis diagnostic tools were be explained. The behavior of each diagnostic under real well determined assessment models and misspecified models was be shown. The usefulness of each diagnostic was be discussed. A new potential diagnostic was be explored (a cohort-based depletion estimator). The recently proposed empirical selectivity diagnostic and its companion R library was mentioned.

The Age-structure production model (ASPM) was proposed by Mauder and Piner (2015) with the goal to understand if the changes in the index of abundance can be explained by the changes in the catches given the fixed selectivities, fixed biology and constant recruitment and assess whether there is enough information in the indices in combination with the catches to estimate the absolute scale of abundance, without the influence of composition data. It is also useful to assess whether there is model misspecification. The ASPMdev variation also estimates the recruitment deviations by fitting only to the index, to understand whether the variability in productivity from year-to-year is needed to estimate the effect of fisheries. This diagnostic has been used in tuna stock assessments. In the NP ALB tuna assessment, the ASPM was useful to evaluate when model misspecification was corrected by redesigning the areas for an areas-as-fleet model. In simulation studies the ASPM showed low Type I error -falsely rejecting a correctly specified model (“self-test”, in Carvalho et al 2017), and reasonable power to detect misspecification in selectivity when is present.

The catch-curve analysis (CCA) was proposed by Carvalho et al 2017. The main goal is to evaluate whether there is information on abundance coming from in the composition data, if this information coincides with the information on abundance coming from the indices, and whether there is model misspecification. If the composition data are driving the IM results the CCA and the IM should have similar results. If the model is correctly specified, trends in abundance will be like those estimated in the ASPM (or ASPM dev). In EPO YFT assessment the CCA was useful to detect the misspecification in selectivity of fleet that was assumed to have asymptotic selectivity became apparent (there was a block in selectivity needed in recent years as larger fish started to dominate the length frequencies of the longline fleet). Simulation studies showed that the CCA may have large type I error – falsely detected model misspecification in the correctly specified model but large power to detect misspecification in selectivity when is present.

Empirical selectivity is the ratio of the number of fish at length (or age) class in the catches to the corresponding numbers at length (or age) in the population, obtained from preliminary fit of the population models that assume simple selectivity functions (such as double-normal or logistic selectivity) (Maunder et al. 2020a; see Minte-Vera et al 2020, Xu et al 2020 for applications). The comparison of the assumed selectivity and the empirical selectivity can elucidate whether the selectivity assumptions are coherent with the observed length (or age) composition data. The diagnostic differs from the residuals in detecting misfit of selectivity functions because it gives more weight to the lower and larger sizes. The empirical_selectivity R library (Olivero-Ramos 2021, available from Github) can be used to explore selectivity functions that would best fit the empirical selectivities externally from the integrated model, speeding the model selection process. It is also useful for investigating which fishery should have the assumption of asymptotic selectivity.

The final diagnostic mentioned was a depletion estimator of abundance at age based on age specific indices of abundance (Clark 2020). The depletion estimator gives multiple estimates of abundance for each cohort at different ages, which are all linked through catch and natural mortality. The population can be represented by a parameter for each cohort and the numbers at age are reconstructed for each cohort. These are then fit to the estimates of abundance at age and time from the depletion estimator to estimate the parameters. The diagnostic can be useful to explore whether there are conditions for estimability of abundance and natural mortality.

[1] remotes::install_github("roliveros-ramos/fks")
remotes::install_github("roliveros-ramos/empirical.selectivity")

R0 likelihood component profile as a diagnostic tool: thumbs up or thumbs down
Huihua Lee, Kevin Piner, Mark Maunder

Profiling over the parameter of the population scale (unfished recruitment R0, or unfished biomass B0) is often used to assess the contribution (based on the likelihood value) of each data component (e.g., abundance indices and size compositions) in the fishery assessment models. Studies over the past decades have used the likelihood profile over scaling parameter as a diagnostic tool to balance different data sources. However, there is no universal agreement on how to prioritize the data. The purpose of this presentation is to review the recent work that used the R0 profile and to summarize the pros and cons of the R0 profile.

Using Diagnostics in r4ss

Ian Taylor –

The widespread use of Stock Synthesis over almost 20 years has led to the development of a large set of diagnostic tools. I will provide a brief overview of some of the diagnostics and how they are accessed, then focus on a few diagnostics that I think would benefit from more frequent use. I will end with some thoughts on making diagnostics for Stock Synthesis more user-friendly and contributing new diagnostics to the collection.

A Guide to Using ss3diags for Model Evaluation

Megumi Oshima, Eric Fletcher, Henning Winker, Massimiliano Cardinale, Laurence Kell, and Felipe Carvalho

Carvalho et al. 2021 presented practical guidelines for implementing contemporary diagnostic tools for integrated assessment models. Accompanying this study, a suite of functions were developed and published as a publicly available R package, `ss3diags`. This package allows users to apply advanced model diagnostics to Stock Synthesis models and promotes a standardized process of model evaluation. Included in the package are functions that assist in evaluating: 1) the goodness-of-fit, 2) retrospective and forecast bias, 3) prediction skill using hindcast cross-validation, and 4) model uncertainty. A joint-index residual plot (`SSplotJABBAres`) and runs test (`SSplotRunstest`) allow the user to visually inspect model residuals from each data source for systematic trends and diagnose model misspecification. To evaluate model consistency, `SSplotRetro` visualizes results from a retrospective analysis by plotting the trajectories of estimated quantities, either spawning stock biomass (SSB) or fishing mortality (F), from each retrospective "peel" and calculates Mohn's rho (ρ_M), to quantify bias and, optionally, forecast bias from one step ahead predictions. The output can provide an indication of a retrospective pattern and direction of bias in the estimated quantities. The `SSplotHCxval` function evaluates accuracy and precision of predictions from a model using hindcast cross-validations of abundance indices and mean length- and composition time series. This function uses the mean absolute scaled error (MASE) to compare the model forecast value to a naïve baseline prediction where a score less than one indicates the model has prediction skill. To assess model uncertainty, `SSdeltaMVLN` implements delta-Multivariate lognormal approximation to generate joint error distributions for key management quantities (SSB/SSB_{MSY} and F/F_{MSY}), which is less computationally and time intensive compared to other methods. `SSplotEnsemble` plots a set of models to compare estimated quantities and respective uncertainties. While individually each diagnostic is not sufficient to evaluate a model, used all together, the functions in `ss3diags` can assess a model's plausibility and suitability in a comprehensive, systematic way.

On automating assessment model diagnostics and the need for simulation testing

Henning Winker, Massimiliano Cardinale , Francesco Masnadi , , Laurence Kell , Iago Mosqueira , Felipe Carvalho

As stock assessment methods become more diverse and complex, the need for best-practice guidelines on model diagnostic criteria is increasingly recognized. For example, ICCAT and IOTC have recommended that objective criteria for model plausibility are used for stock assessments that are intended to provide management advice, and that these criteria shall be based on best practice in using model diagnostics for evaluating (1) model convergence, (2) fits to the data, (3) model consistency and (4) prediction skill, as well as biological plausibility criteria. These four criteria have also started to be used in the benchmark processes of GFCM and ICES for developing reference models, but also for selecting and/or weighting of alternative model scenarios within model ensembles. Often alternative modelling frameworks, such AS integrated assessment models (e.g. Stock Synthesis), statistical catch age model (e.g. a4a, SAM) and biomass dynamic models, are run in parallel, and may even be considered for mixed-model ensemble advice. To effectively implement general guidelines for model diagnostics, it is therefore important that diagnostic tests can be easily run, generalized and transferrable across models. This presentation provides an overview about recent progress in automating selected candidate diagnostics tests for routine use in benchmark assessments of ICCAT, IOTC, GFCM and ICES, with the help of R packages such as `ss3diags`, `JABBA` and `a4adiags`. The introduction of quantifiable diagnostic criteria has generally been welcomed to help with the selection of models for providing advice. However, the sensitivity and specificity of diagnostic tests as well as causes and implications of failing one criteria or the other largely remain open questions. To address these emerging questions, the need and options for simulation testing of model diagnostics are discussed.

Diagnostics, yesterday, today and tomorrow

Andre Punt

Quantitative fisheries assessment has moved from the use of ad hoc model fitting techniques to state-of-the-art approaches. Initially simply being able to estimate parameters and achieving a "reasonable visual fit" was adequate, if not ideal. The concept of multiple model formulations and selecting among models or even weighting them was unknown until the 1980s, although they are now part of the toolbox in many jurisdictions. However, the advent of multiple models has coincided with the need for ways to select among models, including how to decide that a model is sufficiently inadequate not to useful for management / inferences about population trend. Many diagnostics, some based on model fit, others based on the realism of model outputs and the behavior of the model as additional data are added, and others of based on predictive skill have been proposed. However, there is never an ideal approach, and it is recognized that the models used for inference are always gross simplifications of the reality, with the data rarely collected with the primary intent to form the basis for an assessment. This paper will attempt to synthesize the available approaches and (perhaps) provide tentative best practices.