# Evaluating Goodness of Fit

## Goodness of fit is a measure of how well the predictions from a statistical model match the data. A stock assessment model with a good fit to the data is more likely to account for important processes in the population, and to produce useful estimates of stock size, status, productivity, and forecasts to inform management. For a model to be useful, however, characteristics other than goodness of fit should also be considered, e.g. parsimony and out-of-sample predictive ability. Here we describe three approaches/methods used to evaluate the goodness of fit of a model: residuals, likelihood, and posterior predictive checks.

## Residuals

**Goal**: Compare predictions to the data, and produce a model in which residuals are minimized, randomly distributed (display no prominent patterns), consistent with distributional assumptions, have few outliers, and do not display systematic deviations (bias).

**Description:** In statistical models, residuals are the difference between observations and model predictions. Residuals are one way to quantify and visualize a model’s ability to fit the data, or identify lack of fit. Contemporary stock assessment models include multiple data types (e.g. indices of abundance, length and age compositions, catch, discards), and evaluation of residuals is a key component of model development and evaluation.

Many definitions for residuals exist, with the simplest being the ‘raw’ residual, the difference between the observed and expected (predicted) values. Raw residuals can reveal lack of fit such as residual patterns, but extreme values cannot be identified on the basis of residual magnitude alone and interpretation can be difficult for asymmetrical and/or discrete distributions. For these and other reasons, multiple residual definitions have been developed, each with properties that increase their utility in assessing goodness of fit.

Residuals which are scaled by a measure of variability can help identify observations that are unlikely given the model (‘outliers’). A common approach is to divide the raw residuals by an estimate of the residual standard deviation (sometimes referred to as ‘standardized residuals’). This practice generates residuals which describe the distance between observed and predicted values in standard deviation units. Standardized residuals with an absolute value above a certain threshold (e.g. 2 or 3 standard deviations) can be identified as ‘outliers.’

Pearson residuals are often used to evaluate fits to composition data (e.g. ages and lengths). A Pearson residual is the raw residual divided by the square root of the variance function. Therefore, the definition depends on the assumed probability distribution. For example, Pearson residuals for multinomial data are defined in Stock Synthesis as

( Of,i,z – Ef,I,z ) / sqrt( Ef,I,z \* ( 1 – Ef,I,z ) / sfabs( nf,i ) ) [1]

where Of,i,z is the observed proportion for fishery f, observation i, and age (or length) bin z, Ef,i,z is the corresponding model expectation, and nf,i is the sample size for that fishery/observation. The function sfabs() in ADMB is a smoothed (and thus differentiable) version of the absolute value function. Sample sizes in [1] are adjusted in some data weighting procedures (e.g. McAllister and Ianelli 1997, Francis 2011).

Quantile residuals and randomized quantile residuals (Dunn and Smyth 1996, Hartig 2020) are useful for residual analysis of non-normal, skewed, discrete, and/or censored distributions. This approach calculates quantiles from empirical cumulative distribution functions using simulated data from the model. These can be posterior simulations from a Bayesian model, or draws from approximate parametric distributions in a frequentist model.

**How to:**

Most residual analyses for stock assessment use visual inspection of residual plots. Quantitative diagnostics are possible, but in practice are used less often.

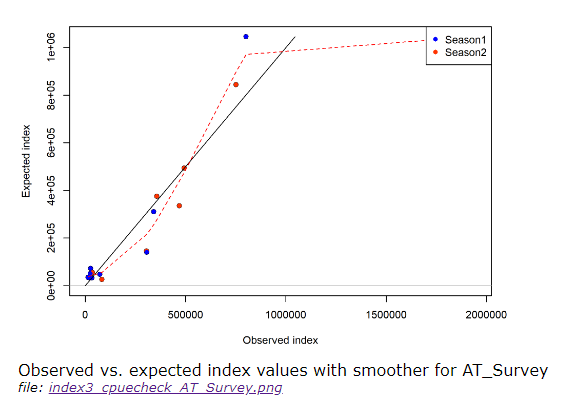
* Examples of residual plots include
  + Observed vs. expected values (residuals implied)
  + Residuals vs. time / fitted values / covariates / omitted variables or datasets
* Characteristics of residual plots to consider:
  + Randomness -- residual patterns, runs tests, correlations (model misspecification), systematic deviations (bias).
  + Consistency with distributional assumptions (mean-variance relationship)
  + Magnitude (predictions within a certain number of standard deviations from the observe values)
  + Outliers and influential observations; observations that are significantly outside the range of model predictions
* Software to automate plot generation
  + r4ss
  + [others]

**What to do:** Add more or less complexity to processes in the stock assessment model depending on the residual patterns. Processes should be informed by knowledge of, for example, life history patterns or fishing fleet behaviors. [work in progress]

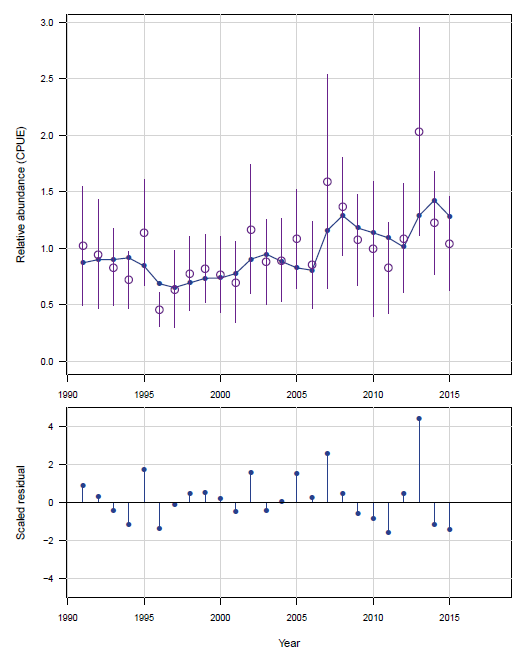
**Examples:**

**SO FAR HAVE MAINLY ‘GOOD’ EXAMPLES; COULD BE USEFUL TO ADD PLOTS WITH ‘ISSUES’**

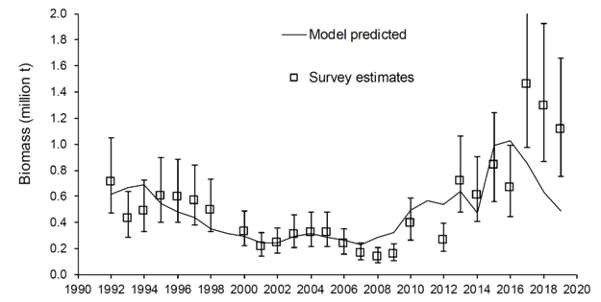
Below is a plot showing observed versus expected index values. Generally, these values are correlated with no major residual patterns.

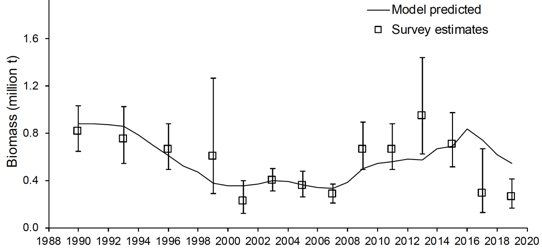


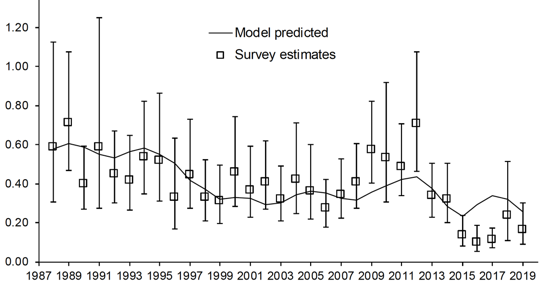
Time series example: Relative abundance (CPUE) time series with observed and expected values (upper panel) and scaled residuals over time (lower panel). Source: SEDAR 2020.



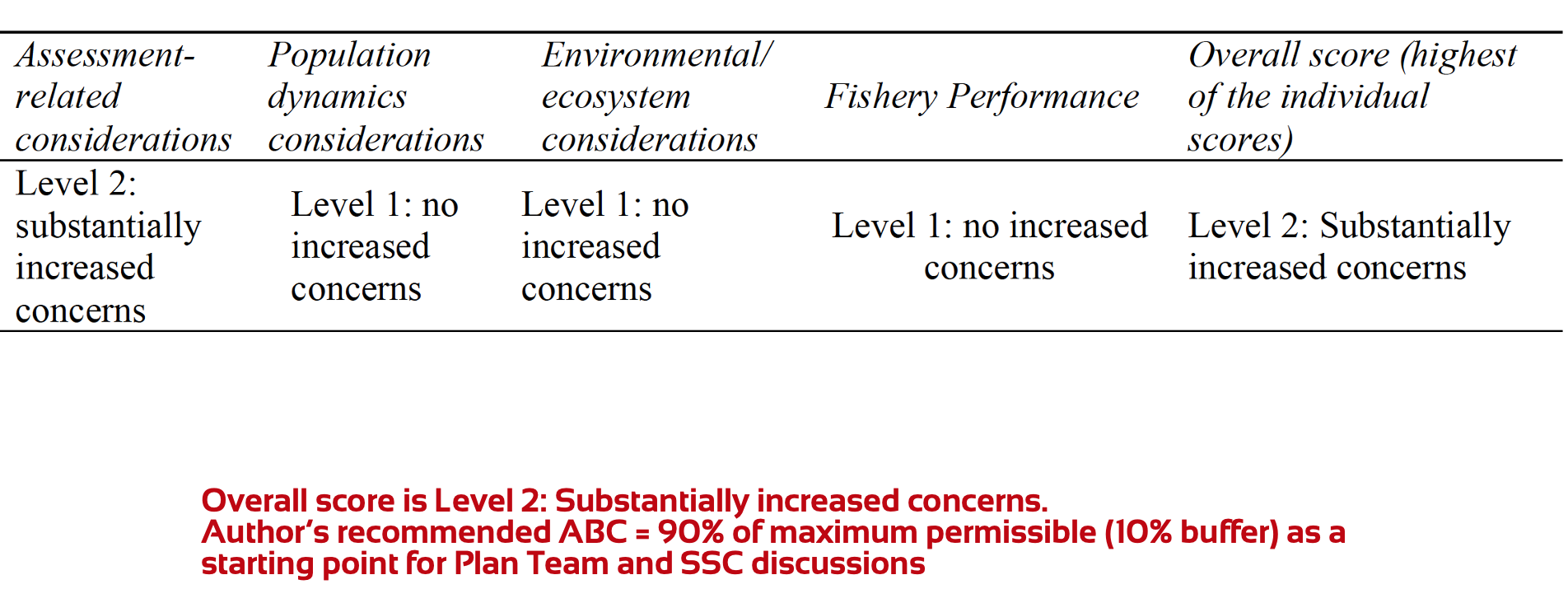
SEDAR. 2020. SEDAR 58 – Atlantic Cobia Stock Assessment Report. SEDAR, North Charleston SC. 500 pp. available online at: http://sedarweb.org/sedar-58







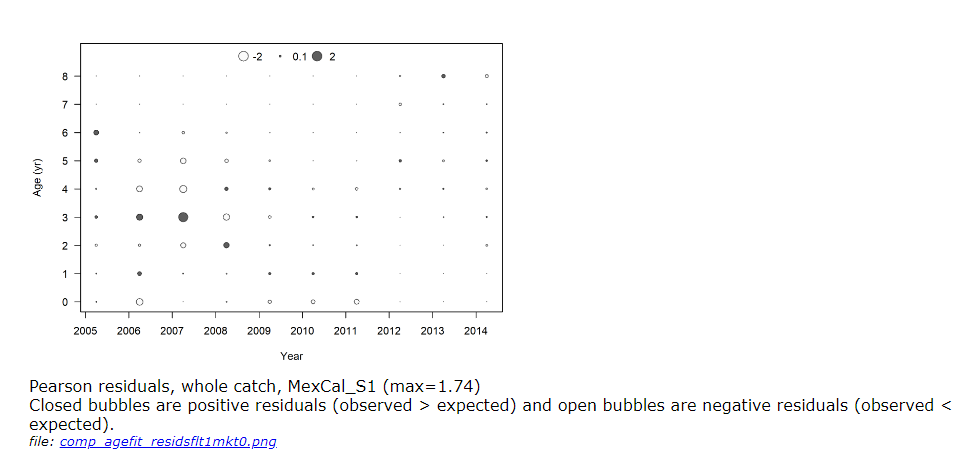
1999 Gulf of Alaska stock assessment model fit to pollock survey data from winter acoustics (top) summer NMFS bottom trawl (middle) and Alaska Department of Fish and Game bottom trawl surveys (bottom).



Assessment concerns affect management advice...

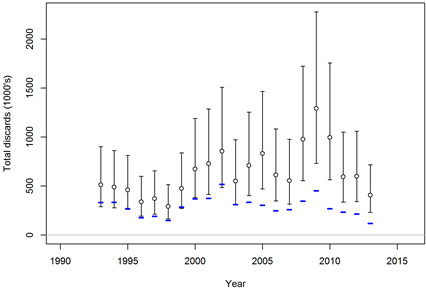
Below is a residual plot for age composition fits from the most recent Pacific sardine benchmark assessment. Overall, fits to the age compositions are good, and there is not much residual pattern. This model has time-varying nonparametric age-based selectivity, which is a flexible functional form without a prespecified shape. Pacific sardine have annual migrations with variable timing and range. The flexible selectivity pattern accounts for year-to-year variability in the fish available to fishing fleets.

A strong residual pattern might be cohorts and years with consistently positive residuals.

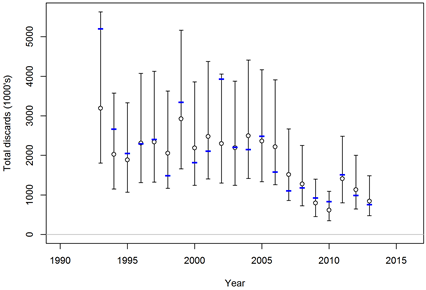
****

**Time series example: Observed (open circles) and predicted discards (blue dashes) in 1000’s of fish of Gulf of Mexico red grouper from the commercial vertical and longline fleets, 1993-2013. Source: SEDAR 2015.**

**Poor fit revealed for the commercial vertical line fishery.**

****

**Better fit revealed for the commercial longline fishery.**

****

**Recommendations:** There are many logistical challenges associated with fishery data. Thus it is unreasonable to expect perfect model fits. It is an analyst’s responsibility to evaluate a range of model configurations and hypotheses to develop a base model with minimal residuals and residual pattern.

## 

## Likelihood

**Goal**: To develop a stock assessment model that minimizes the total likelihood, indicating a good fit to data.

**Description:** Statistical catch-at-age stock assessment models incorporate multiple data types, some which are pre-processed (e.g. standardized index of abundance, age compositions), to characterize the dynamics of a fishery. Comparison of model likelihoods is one means an analyst can compare the goodness of fit of various model configurations.

**How to:** Add more or less complexity to the model to minimize the total likelihood and likelihoods associated with each data source.

* Total
* Components (identify conflicts)
  + By data type (e.g. indices, lengths, ages, etc.)
  + By data source (e.g. fleets within a data type)
  + Anyone use a finer scale? By year, or by data point?
  + “Prior likelihood”
* When it is/isn’t appropriate to compare likelihoods
  + OK if same data, same weights, same ?
  + Adjustments for model dimension
  + Model selection
* Profile confidence intervals
  + R0, h, M, sigmaR
  + 95% = univariate 1.92, bivariate ~3; 0.5\*qchisq(0.95, df=1 or 2)

**What to do:**

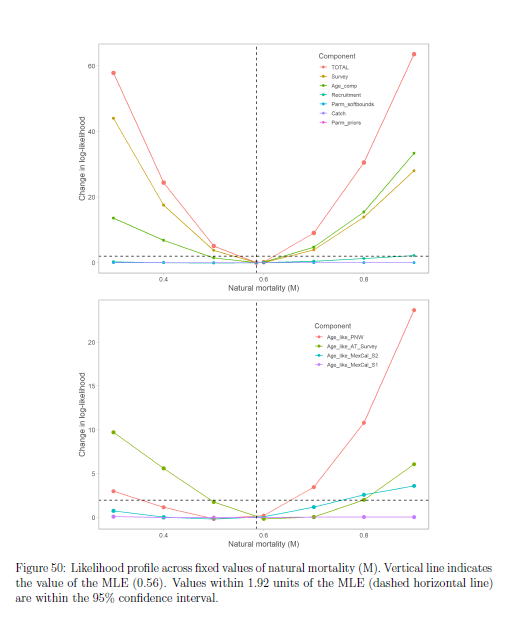
Compare the total likelihood and likelihoods associated with each data source. In these comparisons, an analyst may prioritize one data type more than the other based on prior knowledge.

Likelihood profiles, in which a parameter of interest is fixed and the model is re-estimated, allow the analyst to identify the relative information in each data type and get a sense of the likelihood surface surrounding the maximum likelihood estimates. Likelihood profiles can identify data sets with conflicting information (see Piner plots) and evaluate model sensitivities.

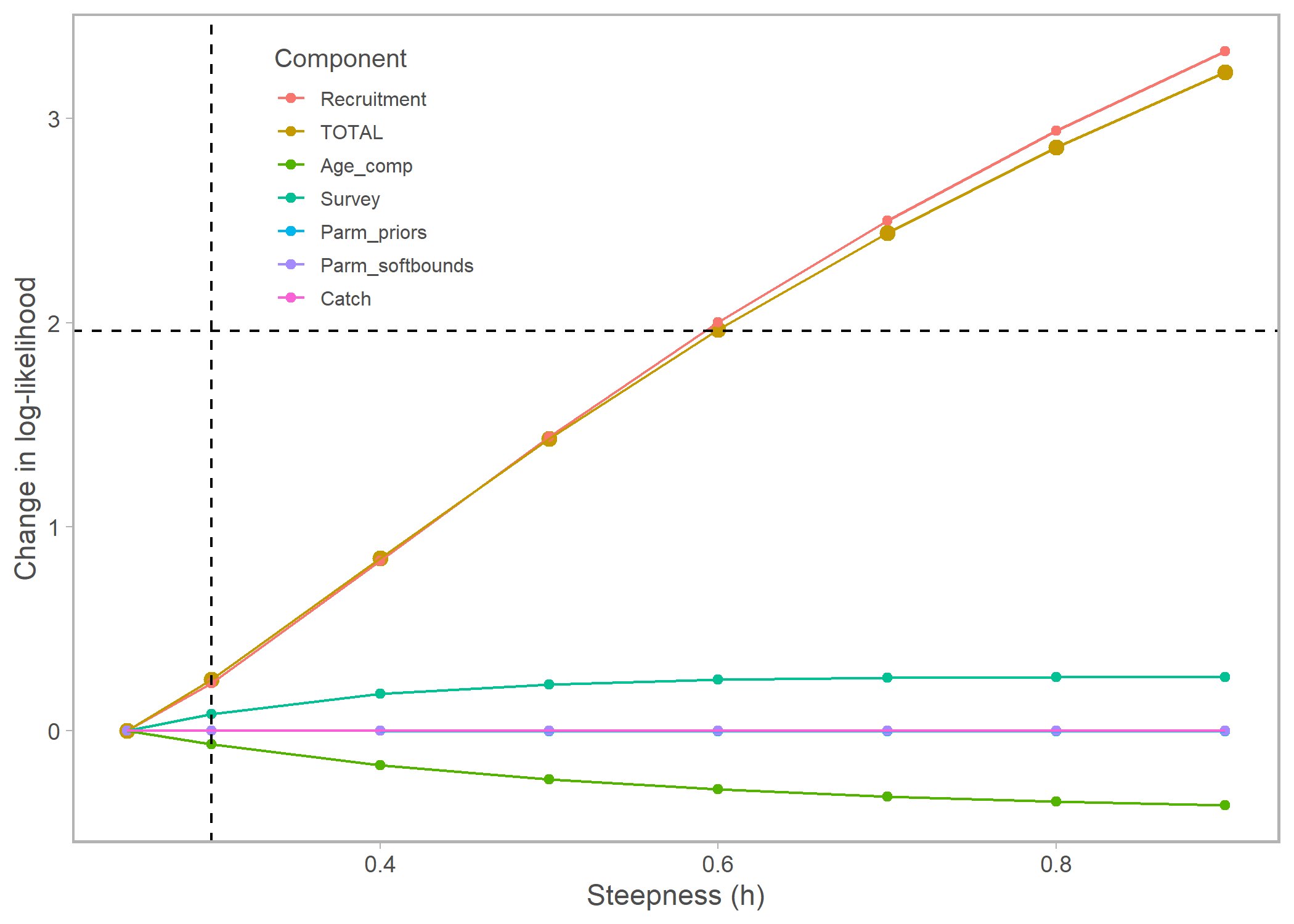
Stock assessment models must balance complexity with the available data, as forecasts are a key component of the fishery management process. Increasing the complexity of a model will increase the number of parameters and likely result in a decrease in likelihoods. Metrics such as Akaike Information Criterion are one means of avoiding overfitting and accurately characterize a population’s dynamics.

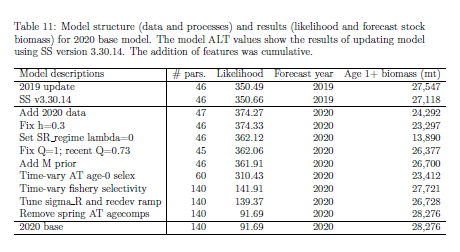
**Example (include figures here):** [more sardine figures because they’re fresh in my mind and easily accessible]

**COULD ALSO USE ‘BAD’ EXAMPLES, REFERENCED IN ‘WHAT TO DO’ SECTION**

****

**Likelihood profile of M**

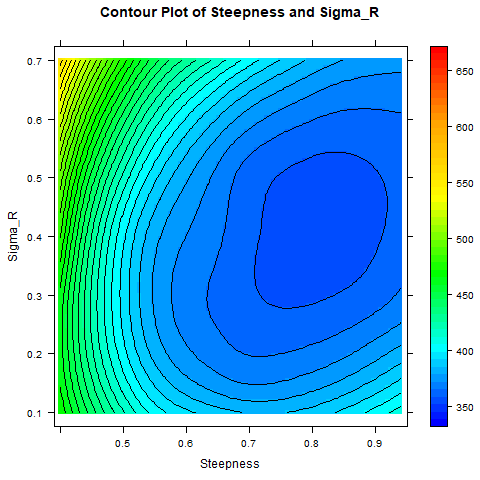
****Piner plot: There is pretty much no information to estimate steepness in the survey or age composition data. This plot justified fixing the value of steepness in the assessment.



Likelihood table showing the likelihoods associated with each step of model development in the 2020 Pacific sardine benchmark assessment.

Bivariate example:

In the Southeast, bivariate likelihood profiles have been carried out where two parameters are fixed across a range of values and the model is rerun for each combination of the fixed parameters. The example below is for Gulf Vermilion Snapper, where profiles were carried out for a combination of steepness and stock-recruit variance parameters. A contour plot, where the z-axis provides the negative log-likelihood value, can then be examined to determine the relationship between the parameters. Although the final model estimates of σR (0.3; eventually fixed at this value in the base model) and steepness (0.71) provide the smallest negative log-likelihood value, a number of alternate pairings give approximately similar negative log-likelihood values. Source: SEDAR 2020 (http://sedarweb.org/sedar-67).



**Recommendations:**

**Key Literature:**

## 

## Posterior Predictive Checks

**Goal**: Evaluate model performance by simulating data sets and comparing to observed data. Compare test statistics applied to simulated and observed data. Useful self-test -- can the model accurately reproduce the data?

**Description:** Simulate data based on the posterior parameter distribution, along with the specified observation error (consistent with specified observation errors).

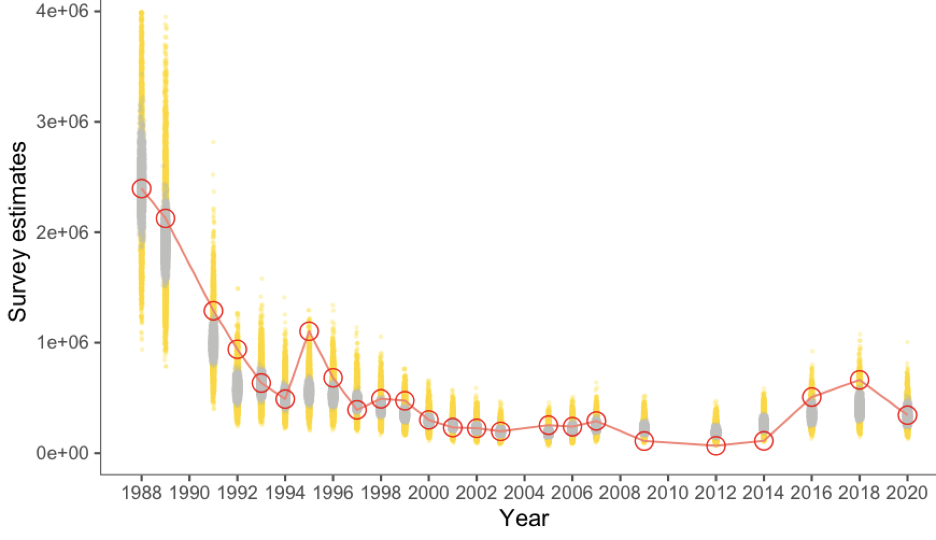
**How to:**

* Simulating data using draws from the posterior distribution
* Could simulate from parametric distributions (asymptotic SEs in non-Bayesian fits)
* Test statistics (mean, SD, min, max, etc.)
  + Aggregated
  + By data source
  + By observation
* Predictive interval coverage

**What to do:**

**Example (include figures here):**

* Mainly biomass dynamics models so far
  + XDB-SRA
  + JABBA (Felipe)
  + SPiCT (https://github.com/DTUAqua/spict)
* Hake?
* Pollock



###### Figure xx. Pollock age-structured model fit to Bogoslof region acoustic-trawl survey biomass estimates, 1988-2020. Open circles and line represent survey point estimates, grey dots represent the posterior mean, and yellow dots represent the posterior predictive values (from Ianelli et al. 2020).

###### 

**Recommendations:**

**Key Literature:**

## Peter’s old comments

**Goal**: To develop a stock assessment model that has sufficient complexity to characterize the major processes in the population without

is unbiased and minimizes residuals (i.e. error).

that fits the data well will

* To develop a stock assessment model that fits the data
  + Model with good fit indicates many of the prominent processes are accounted for and hopefully result in a good prediction
* Evaluate how well model fits data
  + Is there a need to introduce more or less flexibility in the model
  + Similarly what quantitative support is there for assumptions and hypotheses regarding population dynamics

**Description:**

* Caveat that this is biased towards West Coast/Stock synthesis integrated statistical catch-at-age approach (Maunder and Punt 2013; Methot and Wetzel 2013)
* Analysts develop a base model fits well to available data
  + Within joint likelihood framework in which assessment uses data that have been pre-processed (e.g. standardized index of abundance, age compositions)
  + See Maunder and Punt 2013 for description of these models generally and their development through time.

**How to:**

* Quantification of goodness of fit:
  + Likelihoods
    - Total likelihood and likelihoods associated with each data component
  + Parameter uncertainties
    - High standard errors indicate difficulty in estimation (i.e. model may be overparameterized)
  + Effective sample size in compositions
    - The higher the better
* Model comparison (requires specific model configuration and assumptions)
* Visual diagnostics (R4SS)
  + Quick way to see goodness of fit
  + For example, evaluate fits to index of time series (are predictions within confidence intervals of observations?)
  + How well does model fit age or length compositions?
* Avoid residual patterns (also seen through r4ss)
* Model misspecification
* Historical analysis
  + Compare recent model results to past assessment results
* Retrospective pattern (maybe not applicable in this section)
  + Sensitivity to addition of new data
  + Hopefully avoid retrospective patterns, see later section
* Likelihood profiles (also maybe not applicable in goodness of fit section)
  + Quantify sensitivity of MLE
    - Get a sense of likelihood surface
  + Identify parameters for which you may not be able to estimate/identify
    - Flat likelihood profile

**What to do:**

**Example (include figures here):**

Management Councils often require goodness of fit checks as part of the terms of reference.

* [PFMC](https://www.pcouncil.org/documents/2019/04/terms-of-reference-for-the-groundfish-and-coastal-pelagic-species-stock-assessment-review-process-for-2019-2020-april-2019.pdf) examples
  + Appendix B, F2: “Evidence of search for balance between model realism and parsimony. Key model assumptions and structural choices (e.g. asymptotic vs. domed selectivities, constant vs. time-varying selectivities). Summary of alternate model configurations that were examined but rejected.”
* [SEDAR](http://sedarweb.org/sedar-30-terms-reference) examples
  + “Are abundance, exploitation, and biomass estimates reliable, consistent with input data and population biological characteristics, and useful to support status inferences?”
  + “Comment on the degree to which methods used to evaluate uncertainty reflect and capture the significant sources of uncertainty in the population, data sources, and assessment methods.”

**Recommendations:**

See sections below

**Key Literature:**

Maunder and Punt 2013

Methot and Wetzel 2013